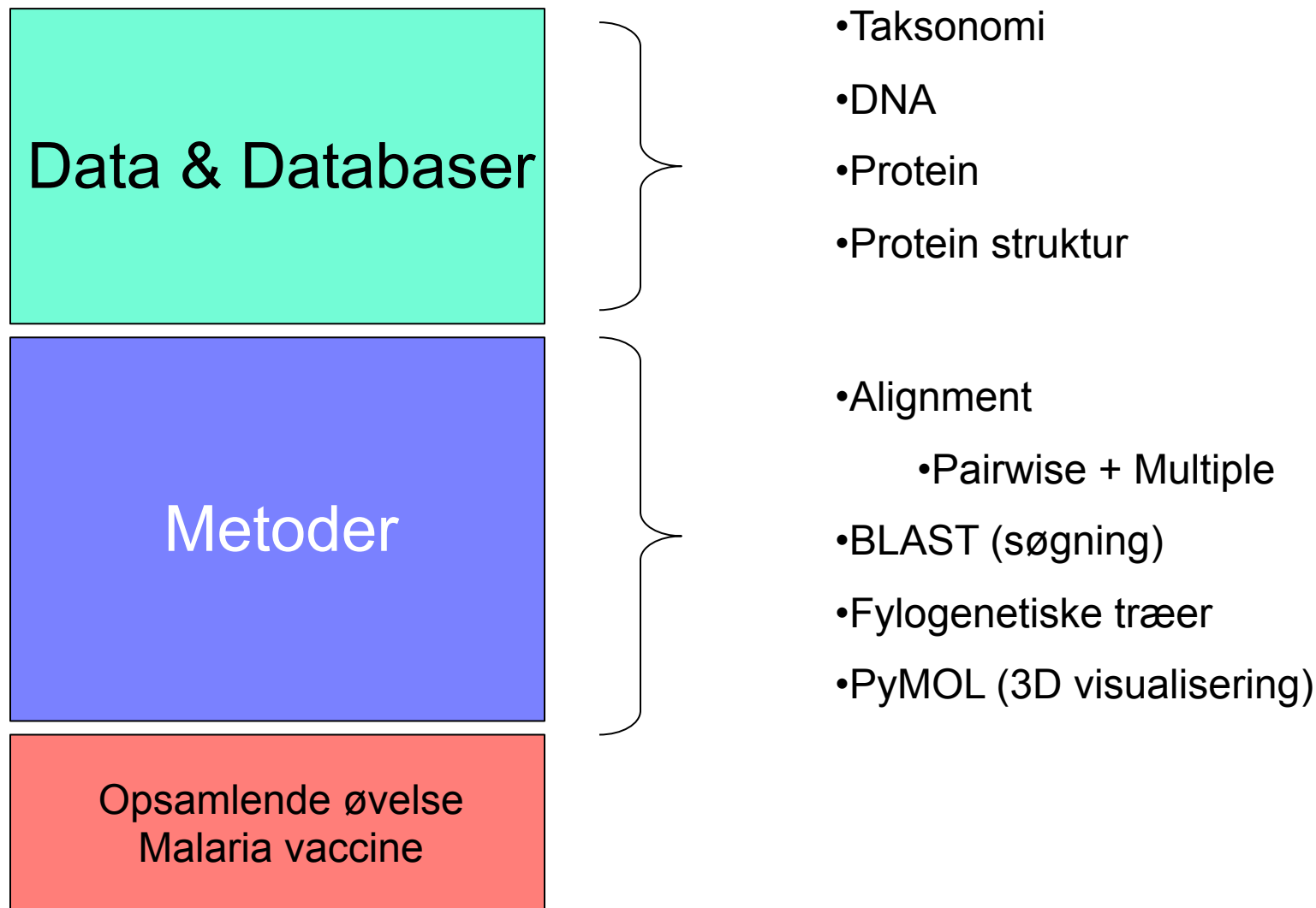
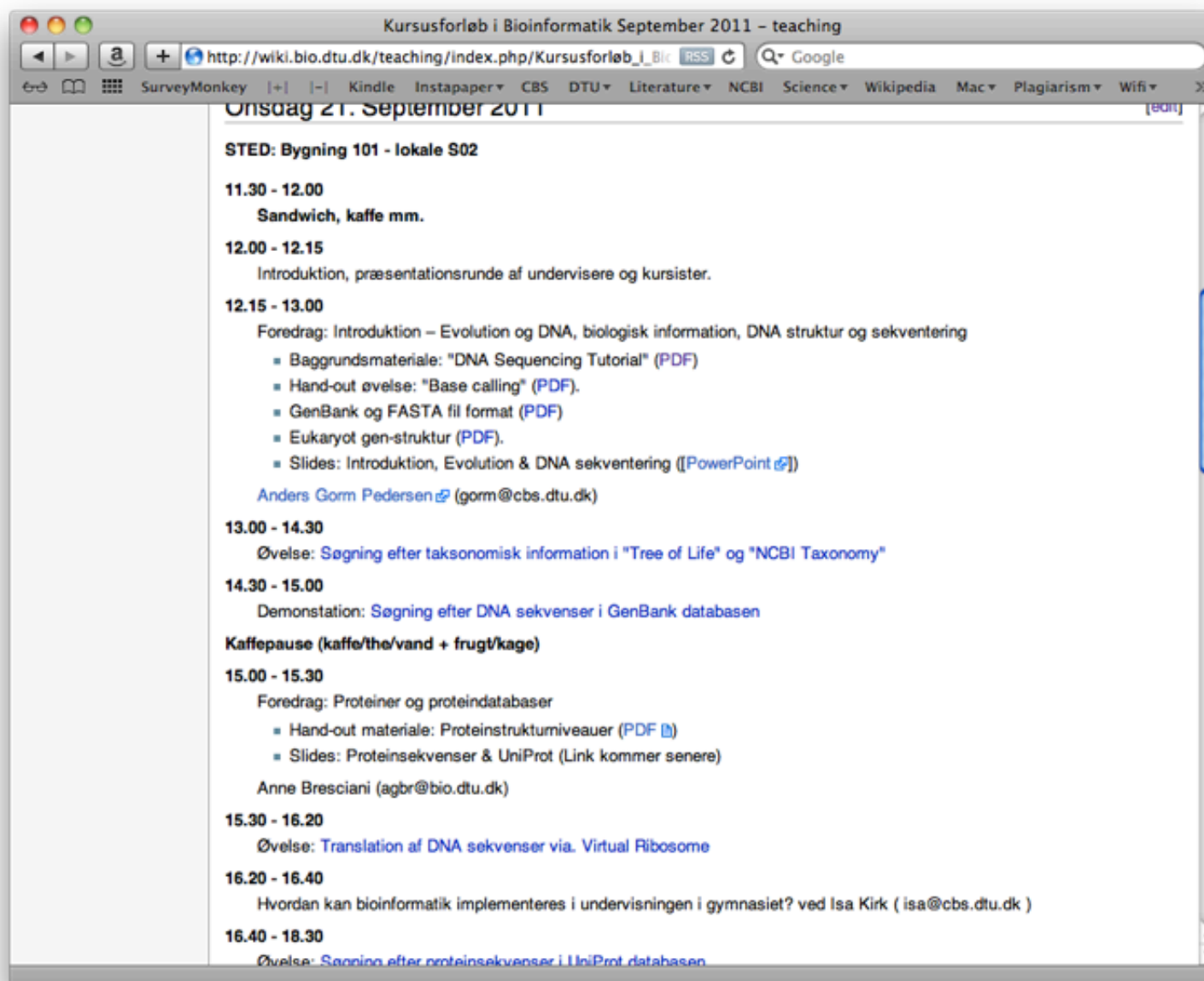


Introduktion til Bioinformatik

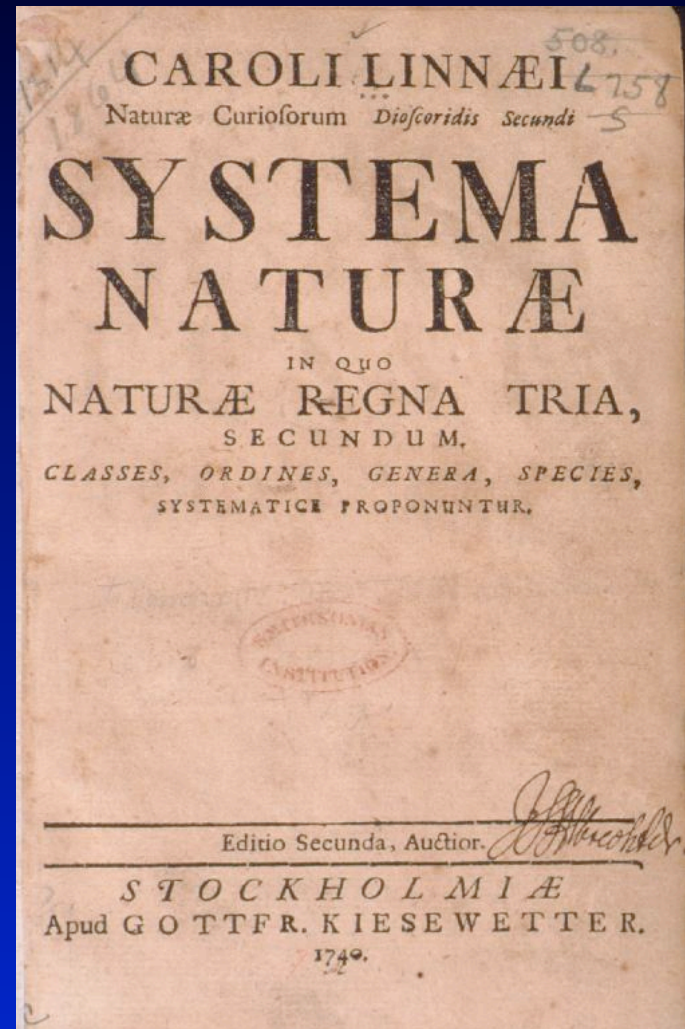




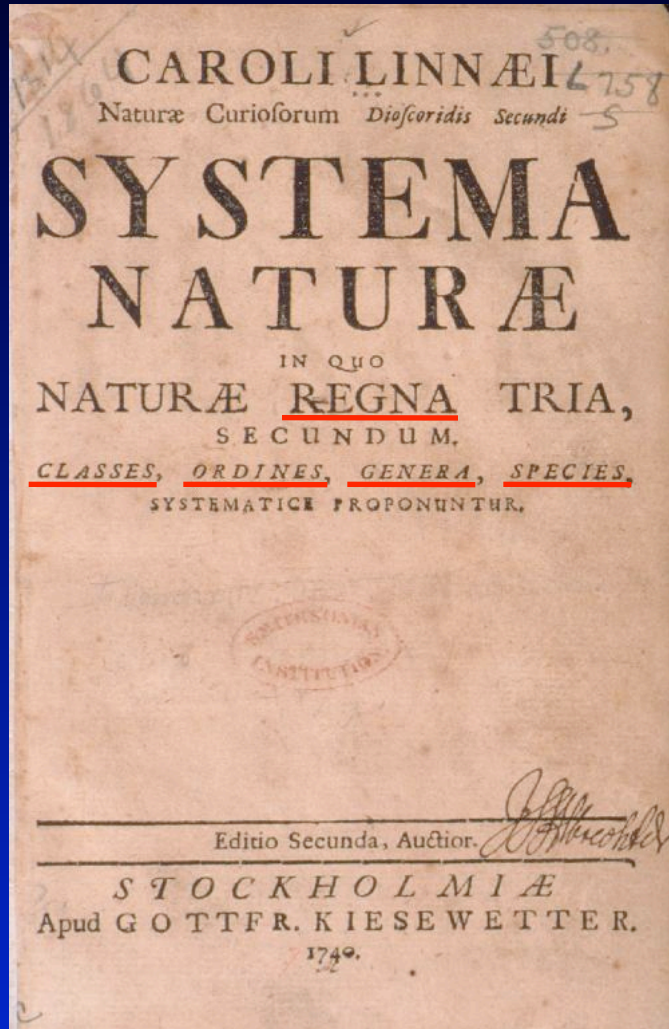
Classification: Linnaeus



Carl Linnaeus
1707-1778

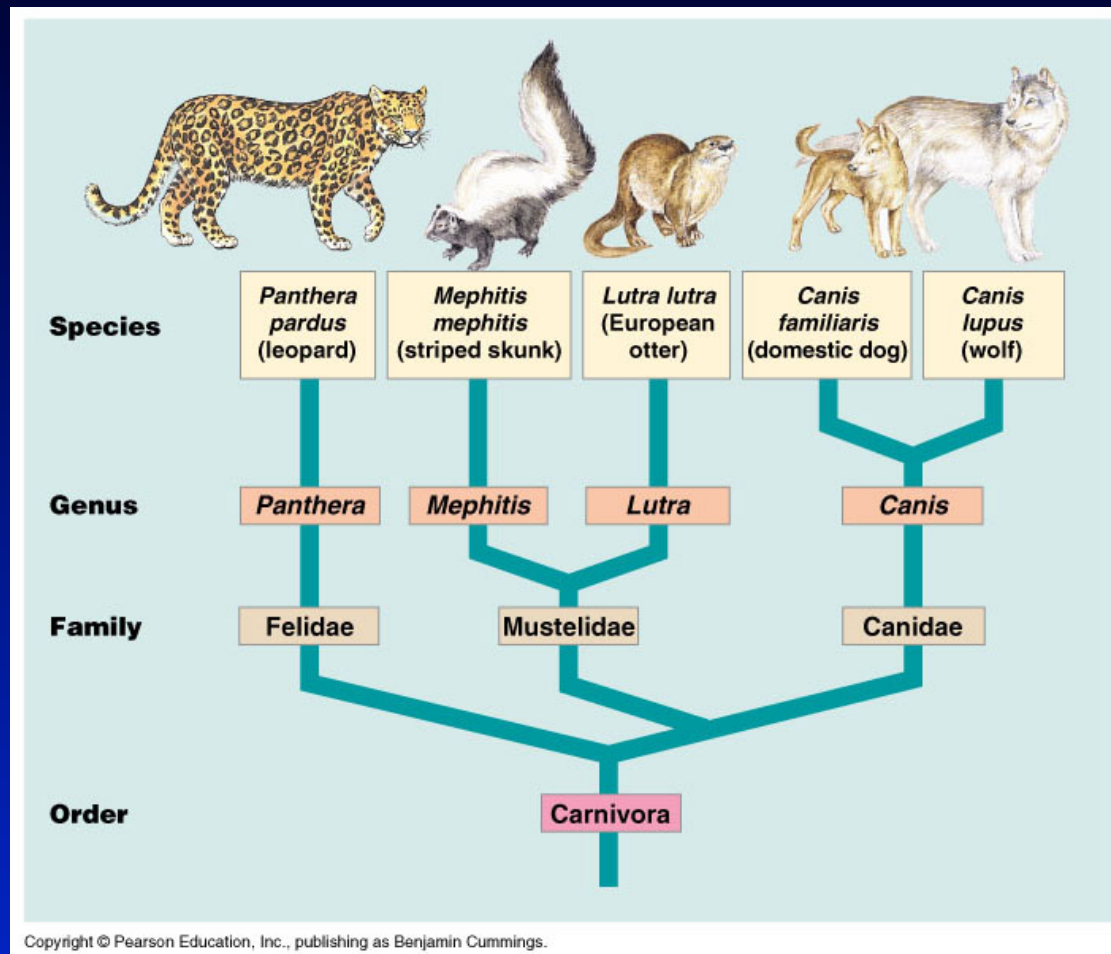


Classification: Linnaeus

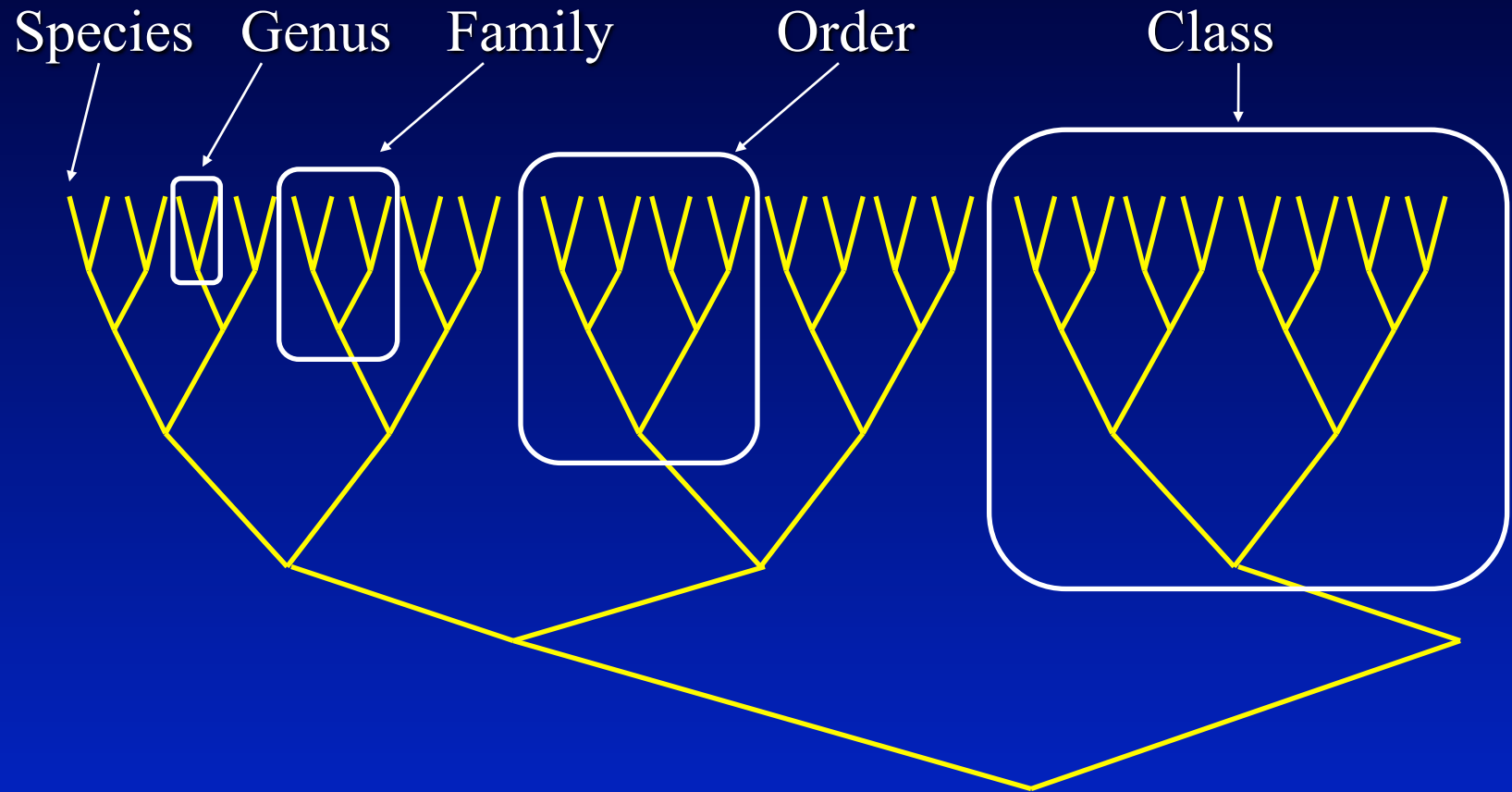


- Hierarchical system
 - Kingdom (Rige)
 - Phylum (Række)
 - Class (Klasse)
 - Order (Orden)
 - Family (Familie)
 - Genus (Slægt)
 - Species (Art)

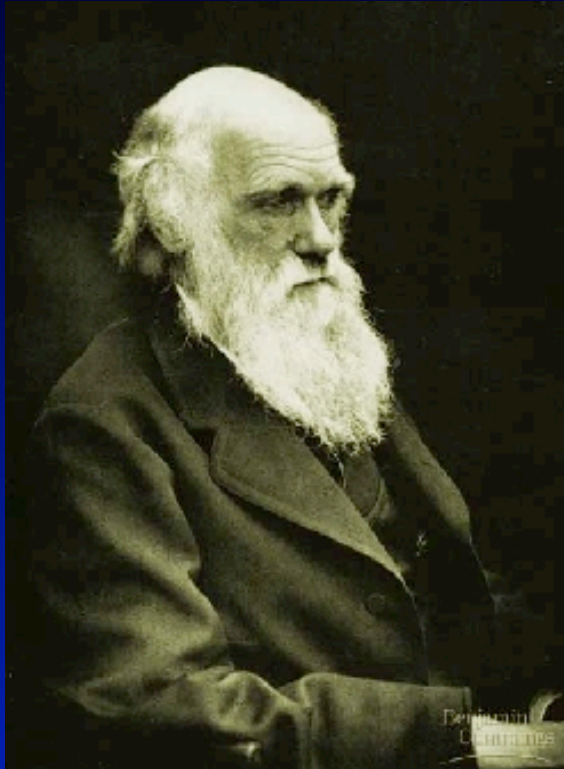
Classification depicted as a tree



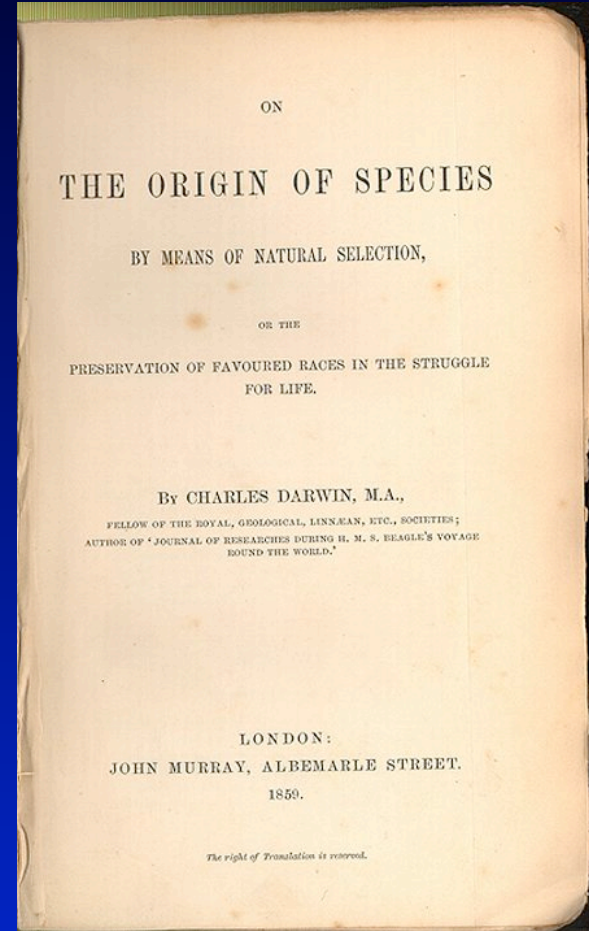
Classification depicted as a tree



Theory of evolution

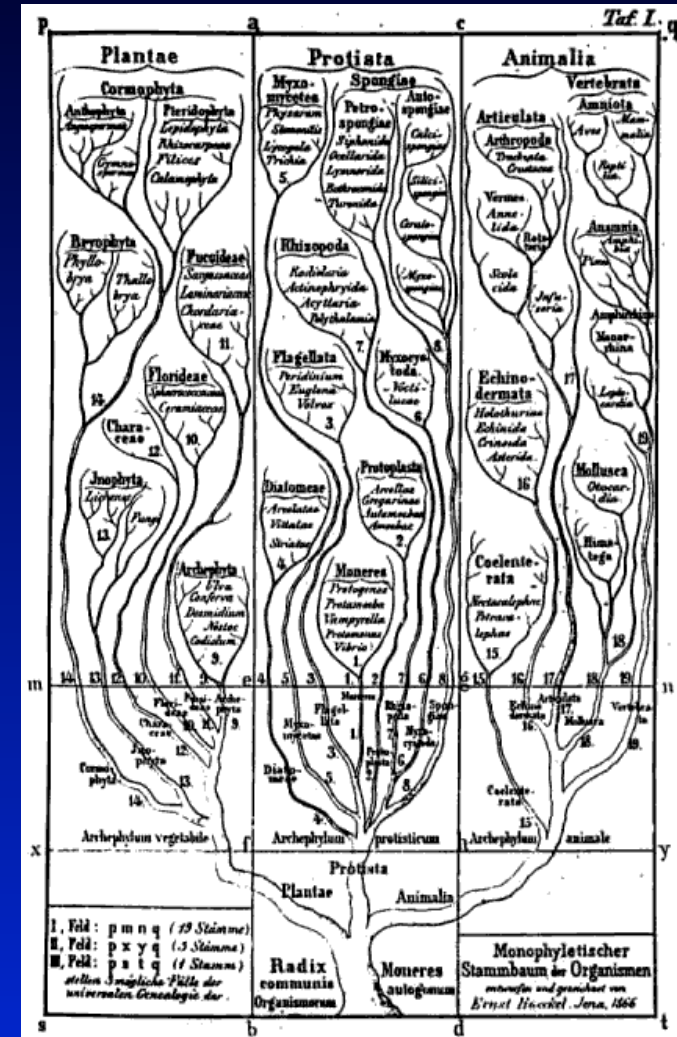


Charles Darwin
1809-1882



Phylogenetic basis of systematics

- **Linnaeus:**
Ordering principle is God.
- **Darwin:**
Ordering principle is shared descent from common ancestors.
- Today, systematics is explicitly based on phylogeny.



Natural Selection: Darwin's four postulates

Natural Selection: Darwin's four postulates

- More young are produced each generation than can survive to reproduce.

Natural Selection: Darwin's four postulates

- More young are produced each generation than can survive to reproduce.
- Individuals in a population vary in their characteristics.

Natural Selection: Darwin's four postulates

- More young are produced each generation than can survive to reproduce.
- Individuals in a population vary in their characteristics.
- Some differences among individuals are based on genetic differences.

Natural Selection: Darwin's four postulates

- More young are produced each generation than can survive to reproduce.
- Individuals in a population vary in their characteristics.
- Some differences among individuals are based on genetic differences.
- Individuals with favorable characteristics have higher rates of survival and reproduction.

Natural Selection: Darwin's four postulates

- More young are produced each generation than can survive to reproduce.
- Individuals in a population vary in their characteristics.
- Some differences among individuals are based on genetic differences.
- Individuals with favorable characteristics have higher rates of survival and reproduction.
- Evolution by means of natural selection

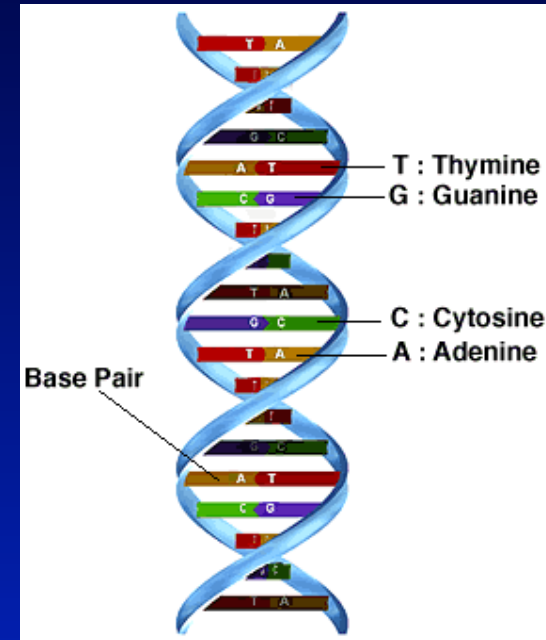
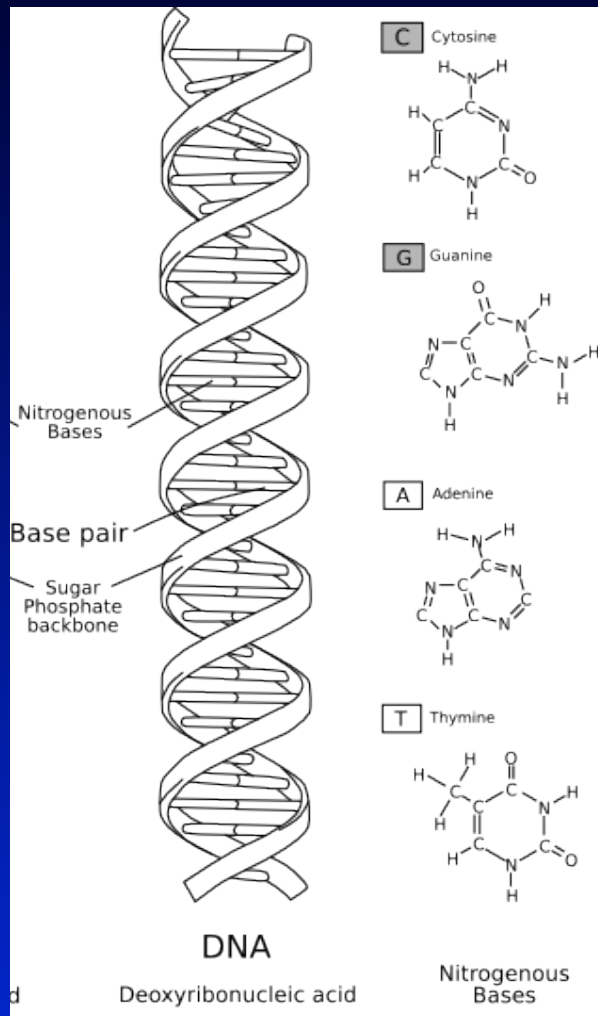
Natural Selection: Darwin's four postulates

- More young are produced each generation than can survive to reproduce.
- Individuals in a population vary in their characteristics.
- Some differences among individuals are based on genetic differences.
- Individuals with favorable characteristics have higher rates of survival and reproduction.
- Evolution by means of natural selection
- Presence of "design-like" features in organisms:

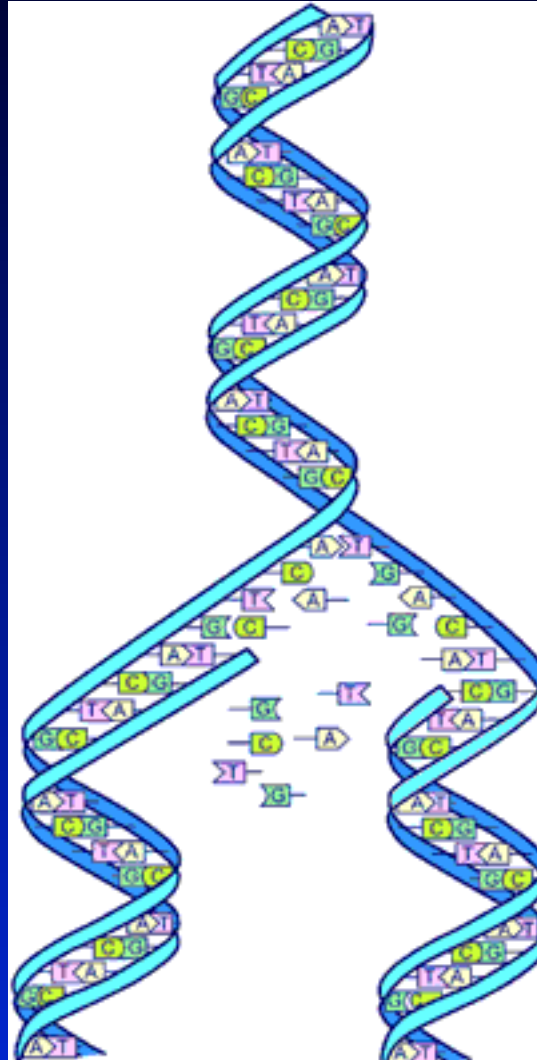
Natural Selection: Darwin's four postulates

- More young are produced each generation than can survive to reproduce.
- Individuals in a population vary in their characteristics.
- Some differences among individuals are based on genetic differences.
- Individuals with favorable characteristics have higher rates of survival and reproduction.
- Evolution by means of natural selection
- Presence of "design-like" features in organisms:
- Quite often features are there "for a reason"

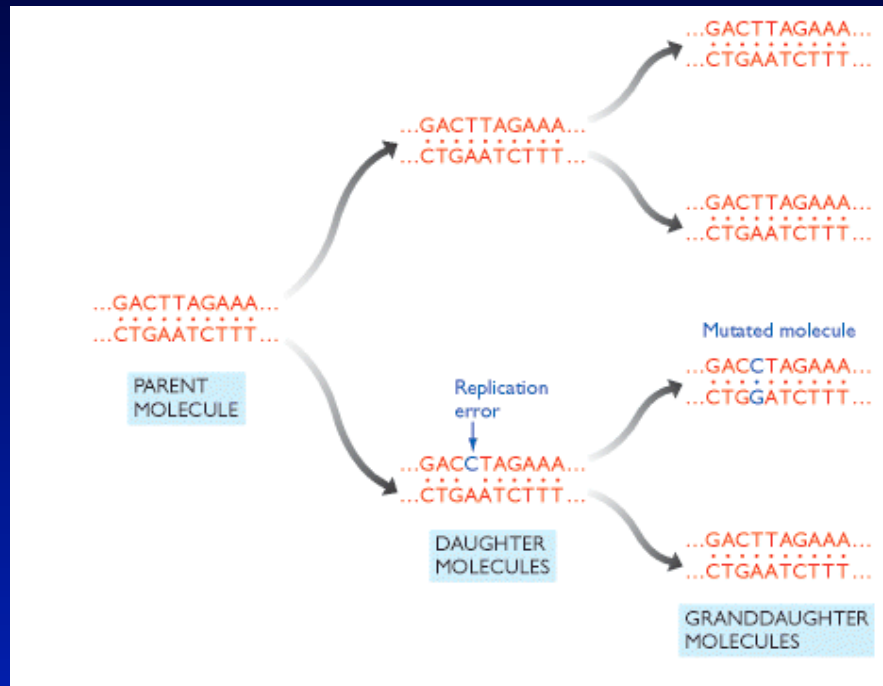
Molecular Basis for Heredity: DNA



Molecular Basis for Heredity: DNA



Molecular Basis for Variation: DNA Mutation



A history of mutations



Time

ATGGCCCTGTGTATGCG

A history of mutations



Time

ATGGCCCTGTGTATGCG

A history of mutations

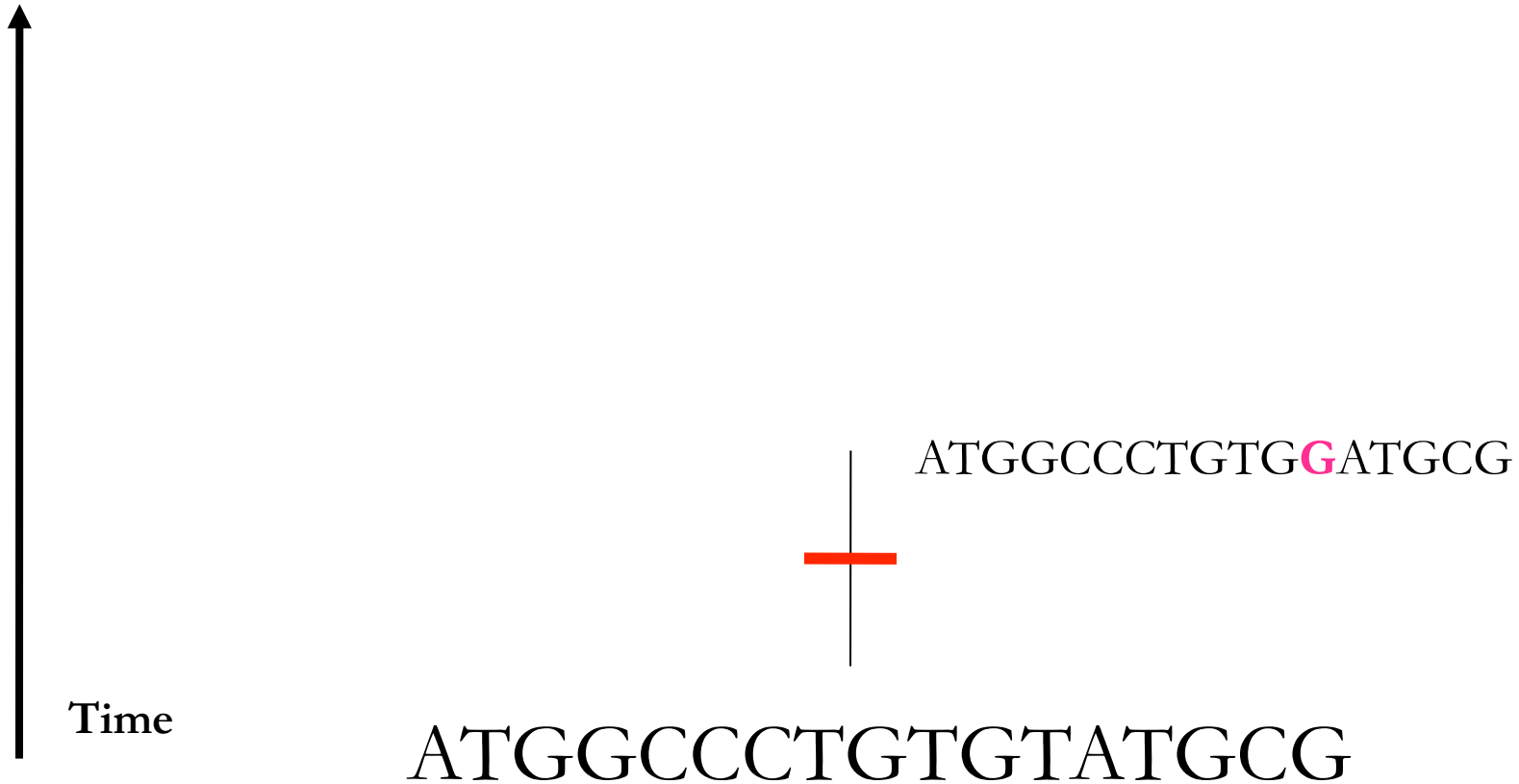


Time

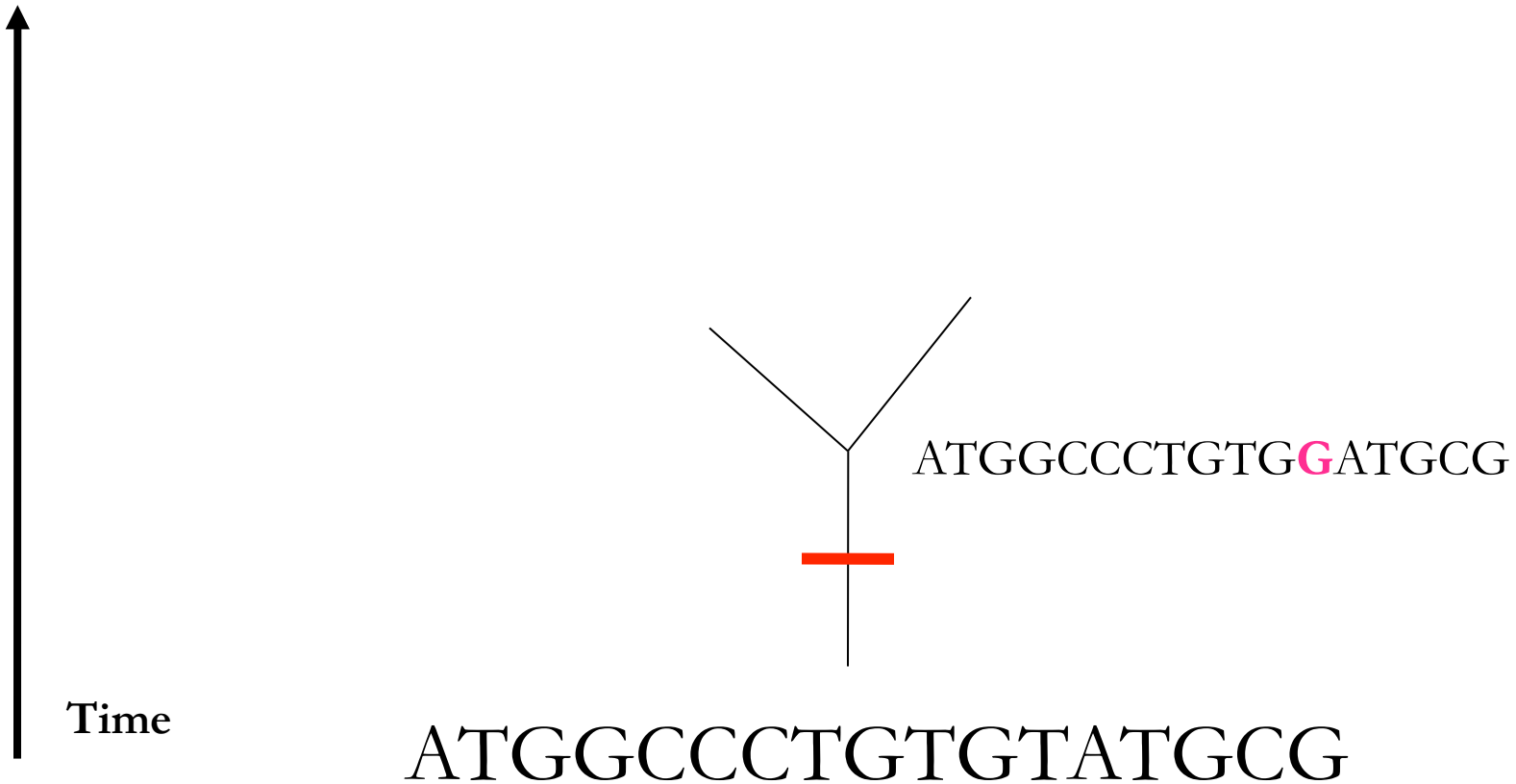


ATGGCCCTGTGTATGCG

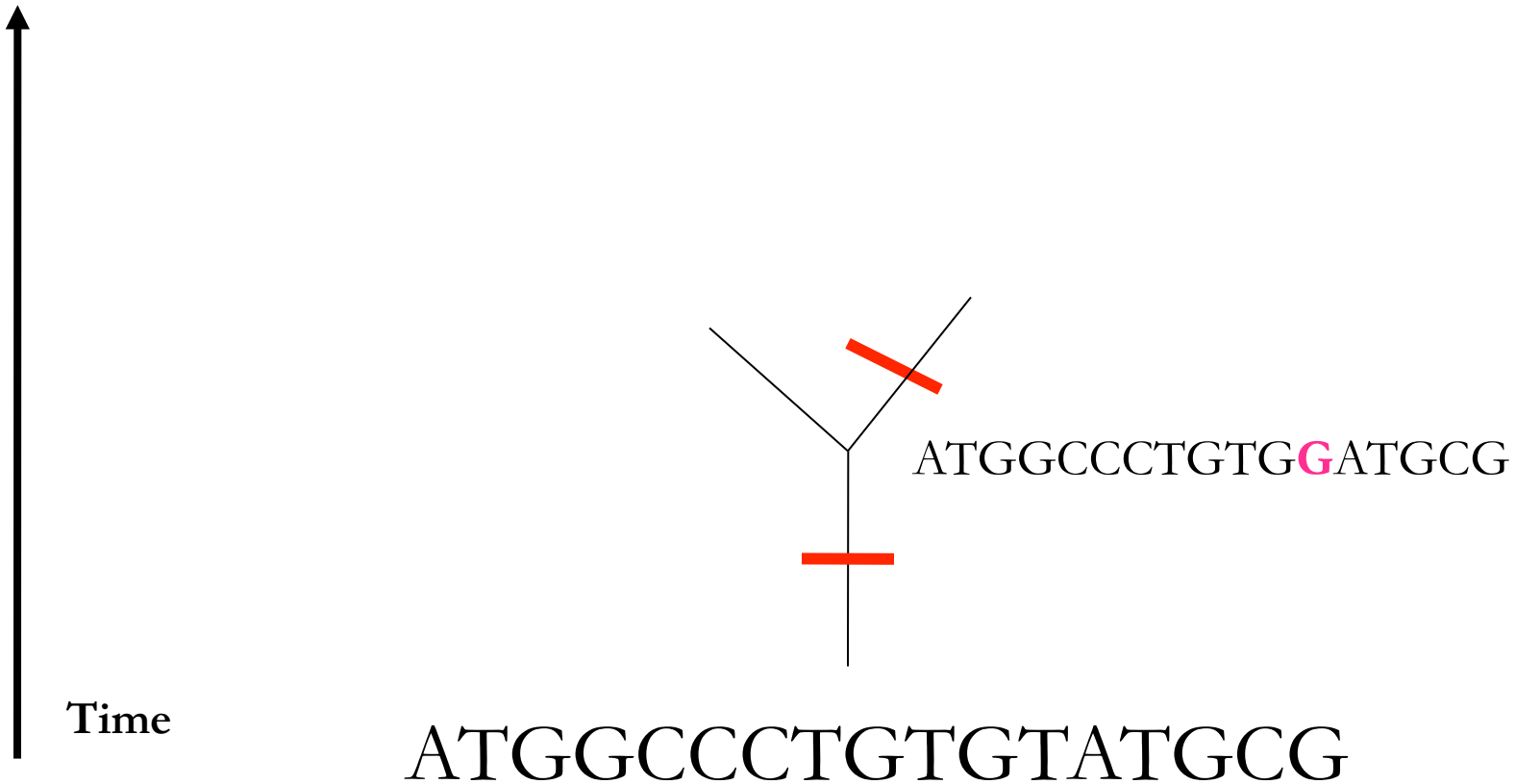
A history of mutations



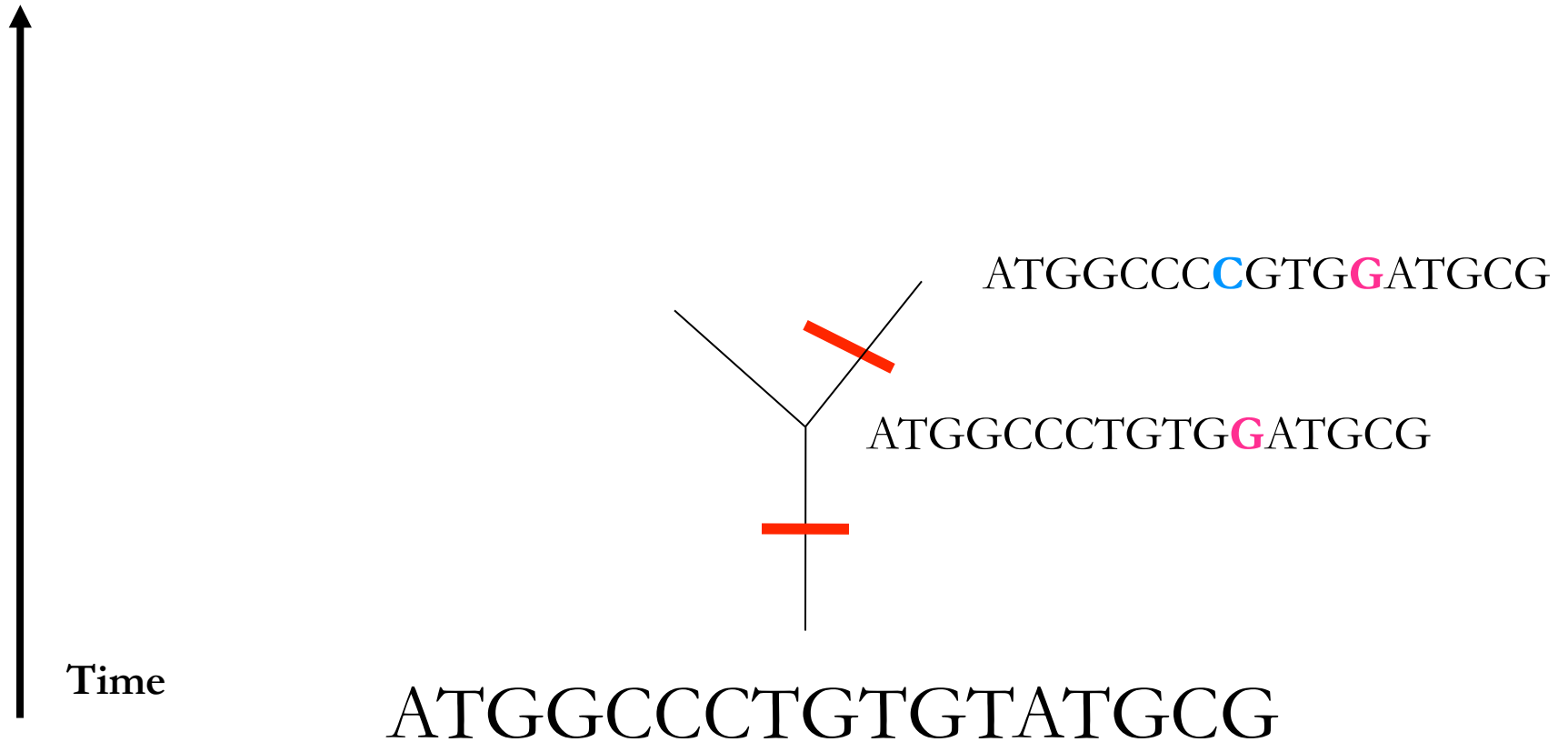
A history of mutations



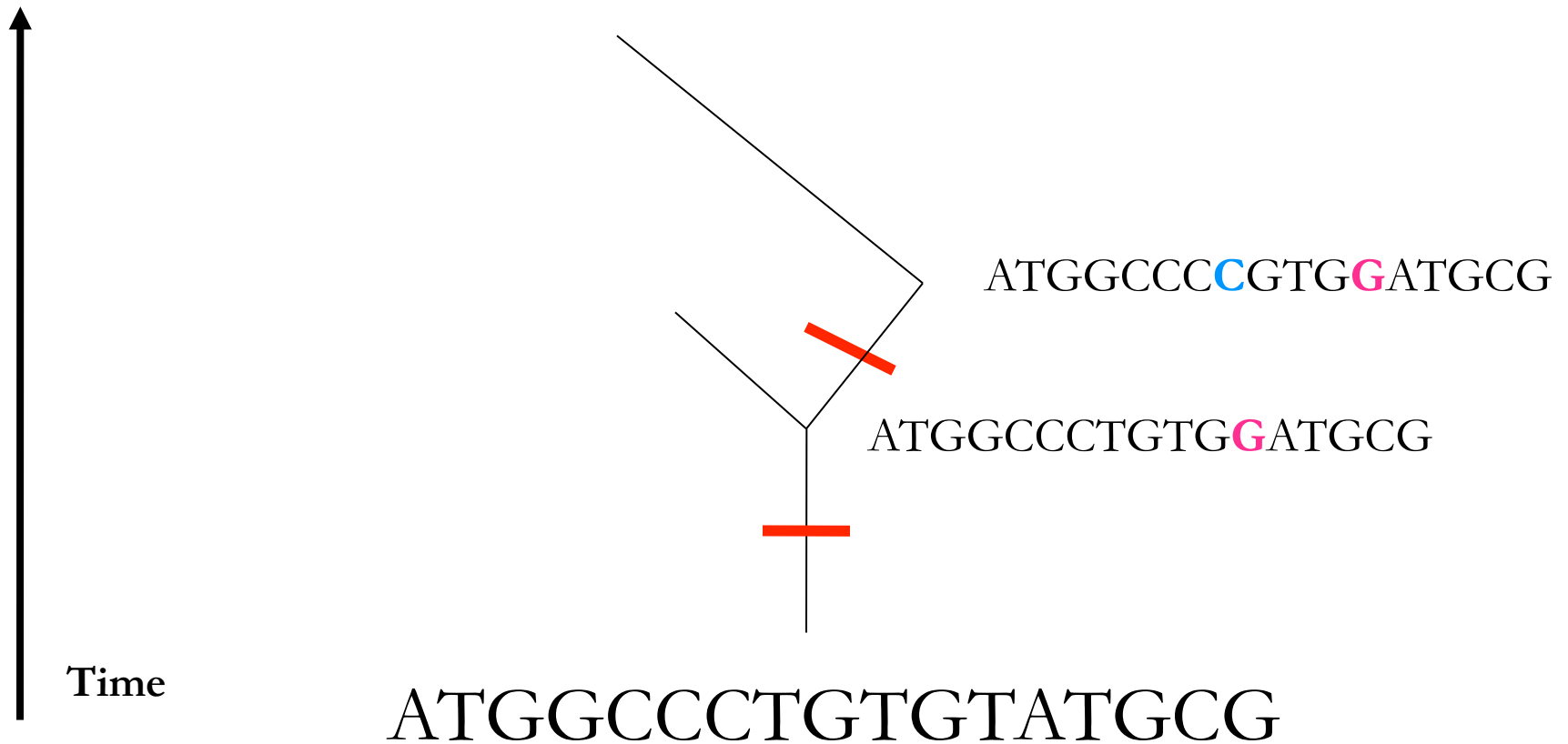
A history of mutations



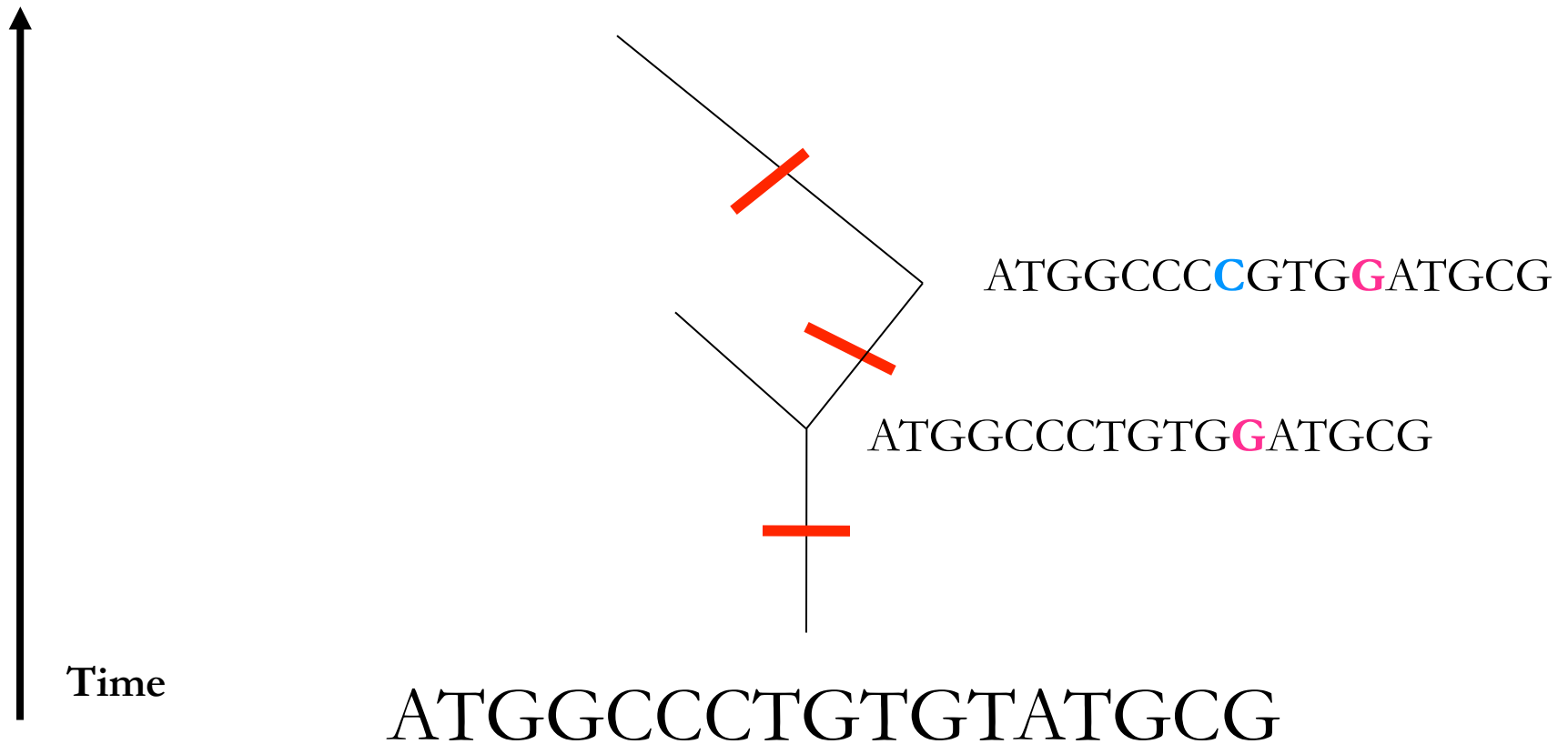
A history of mutations



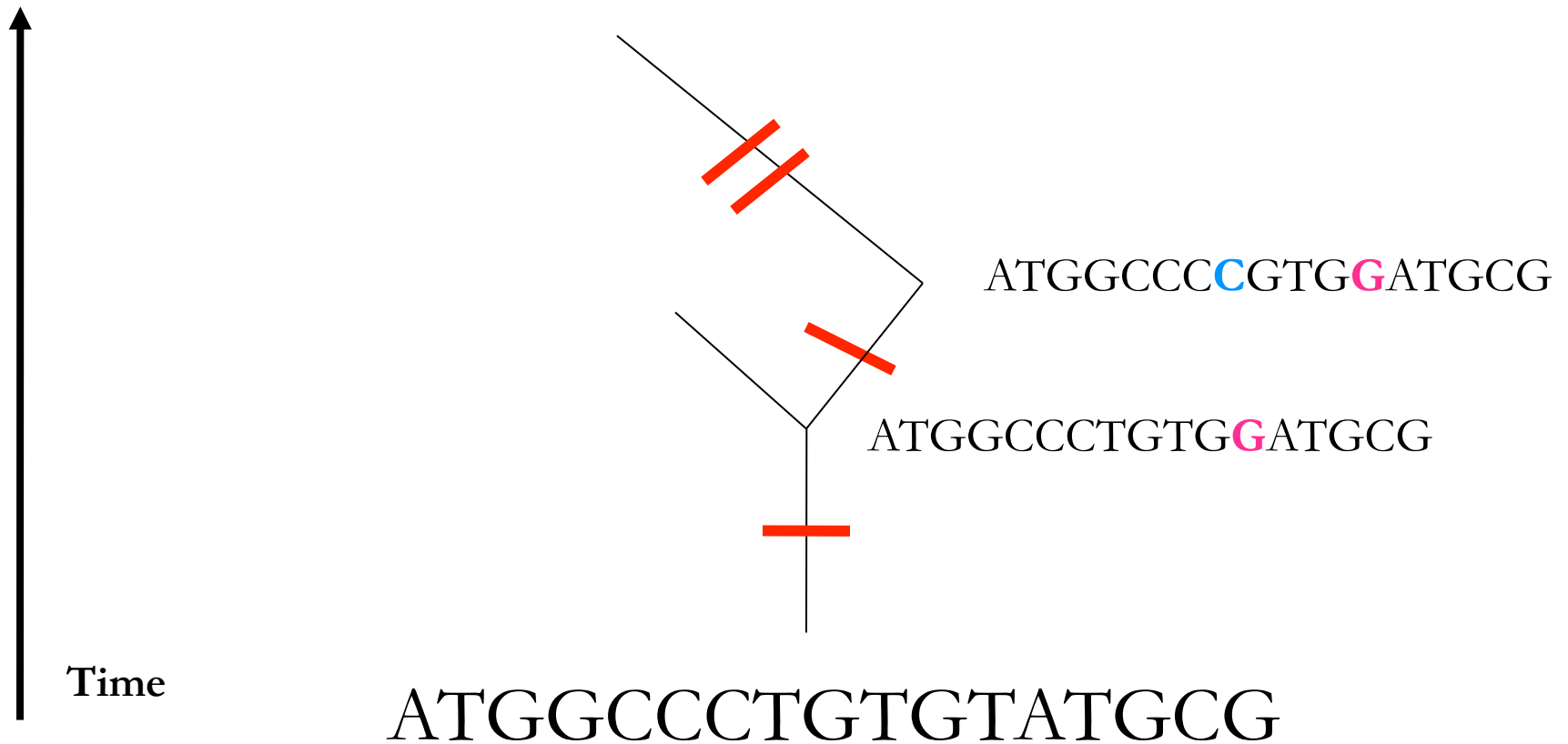
A history of mutations



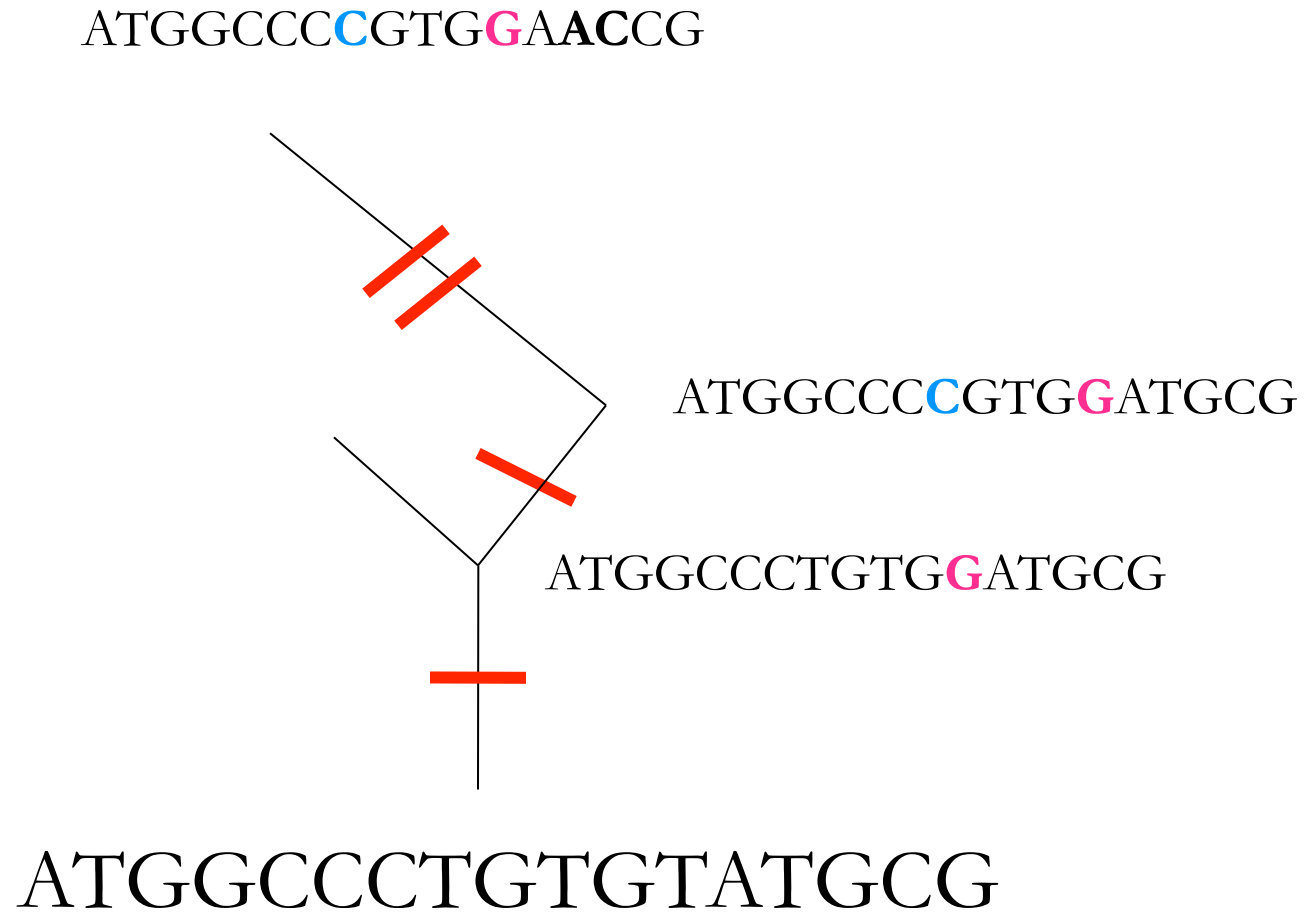
A history of mutations



A history of mutations

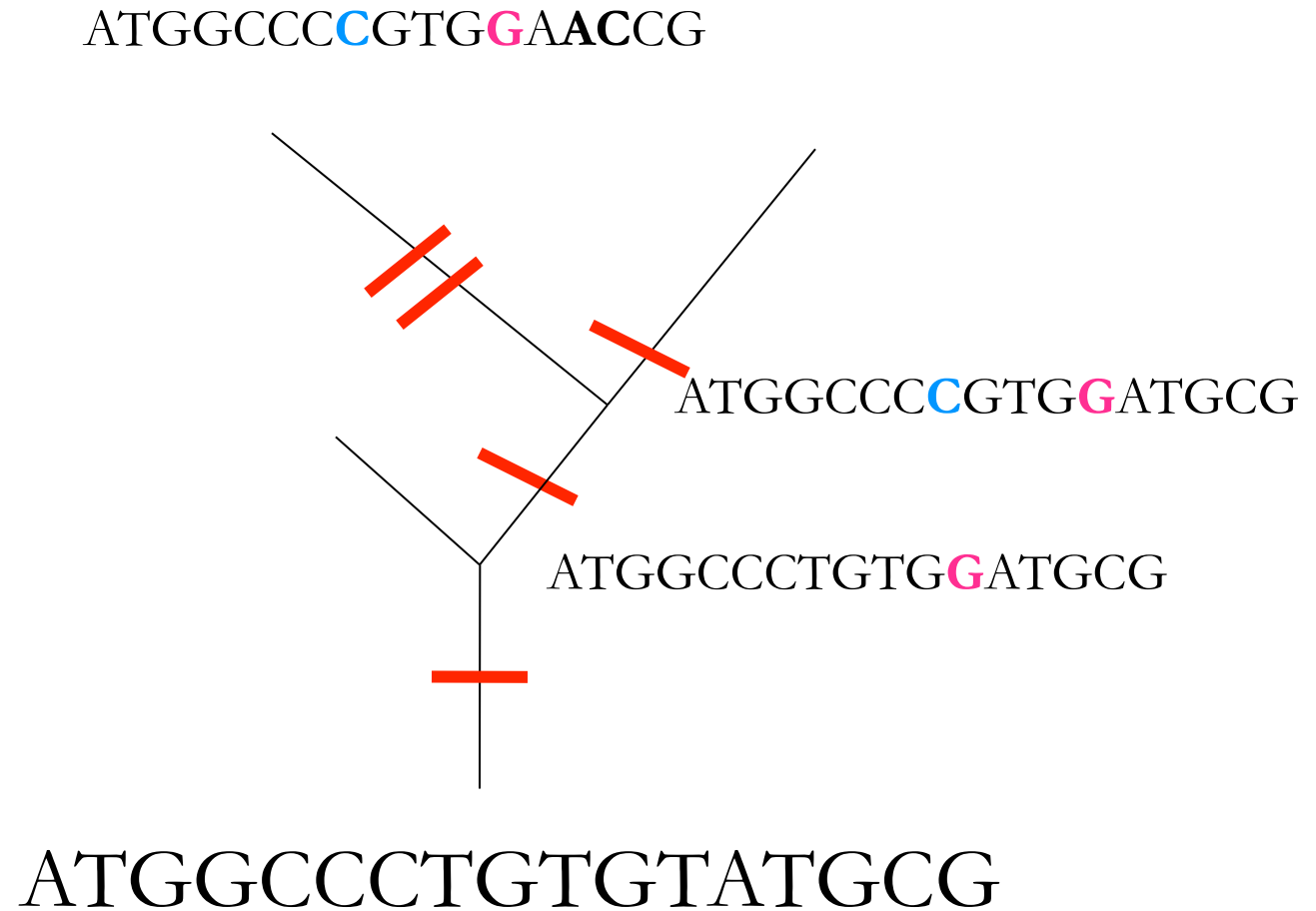


A history of mutations

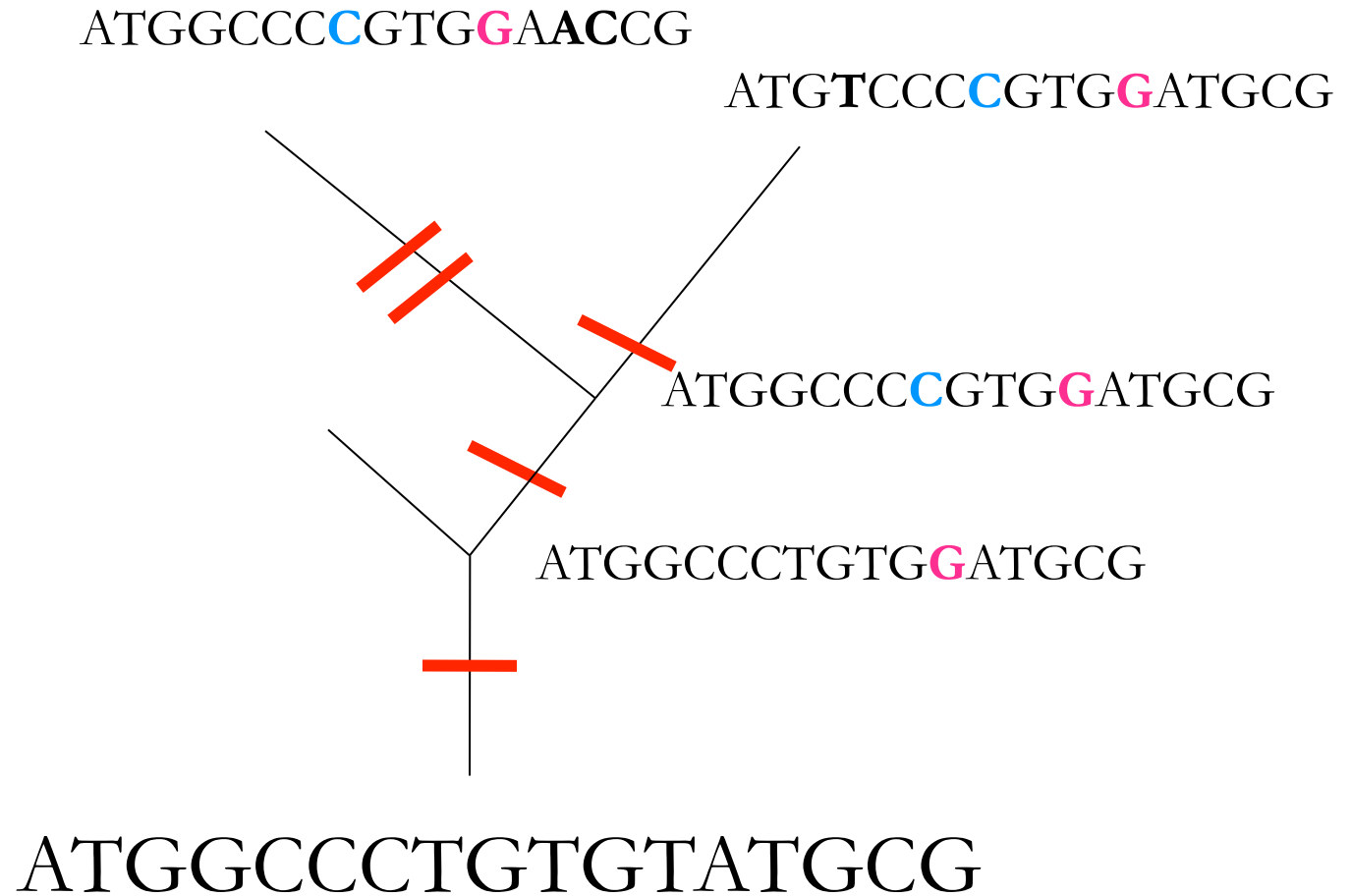




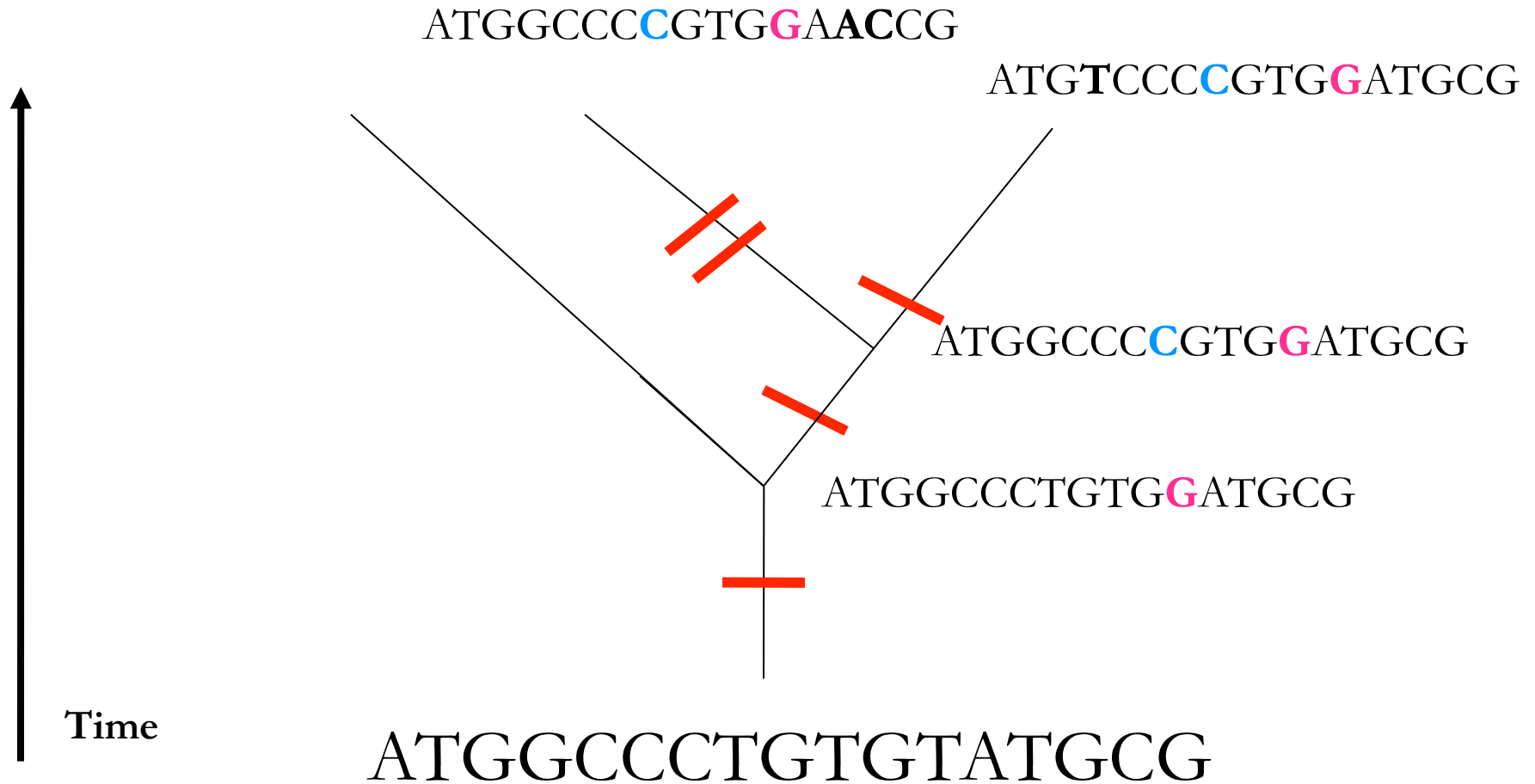
A history of mutations



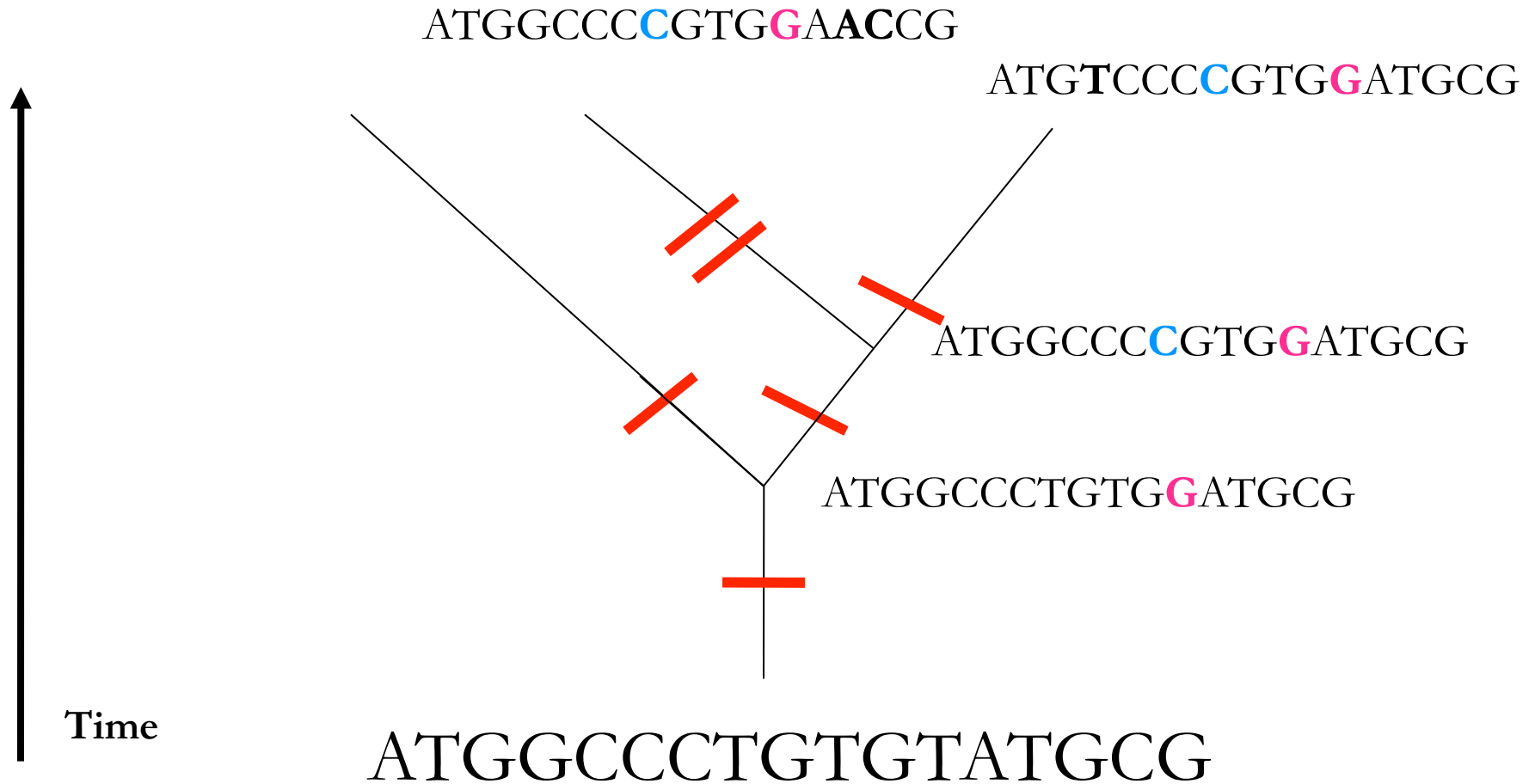
A history of mutations



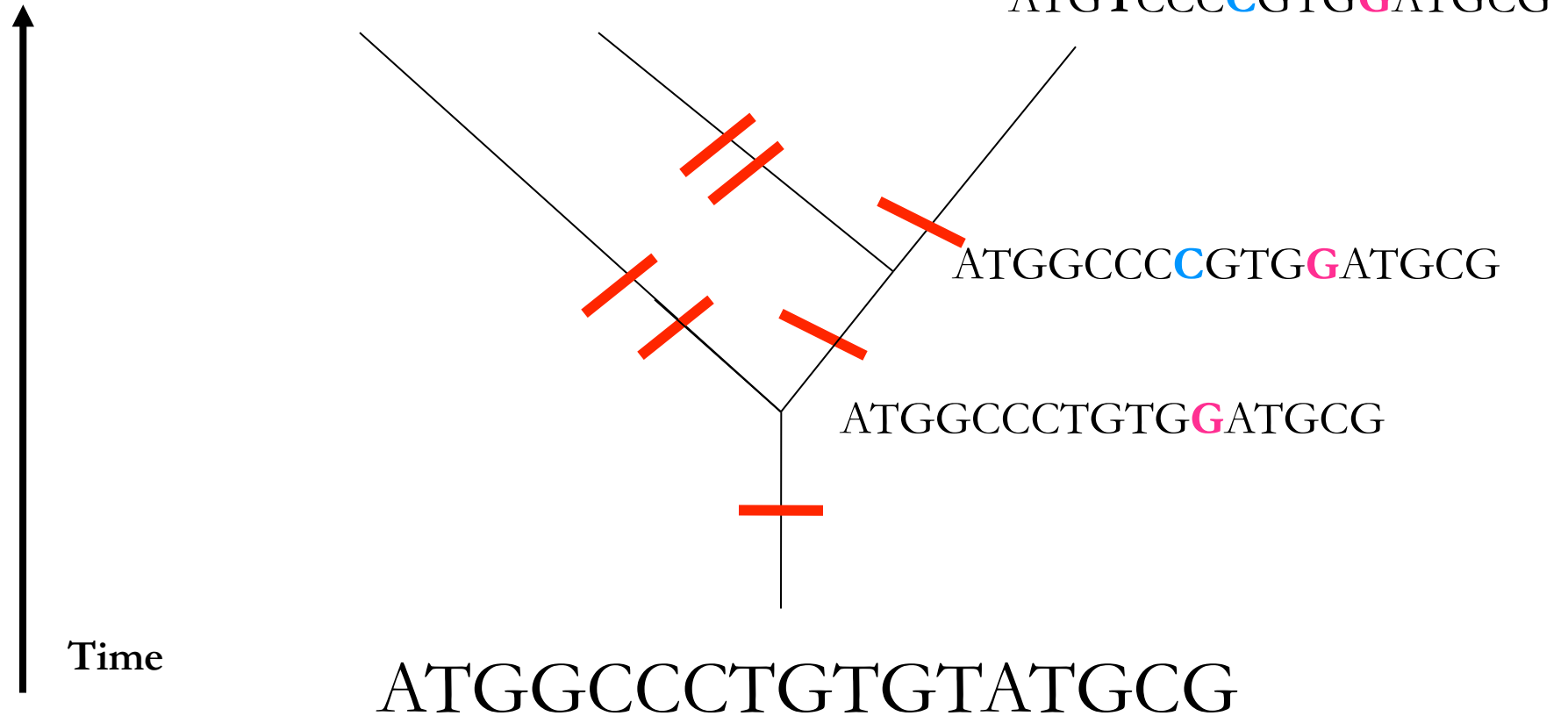
A history of mutations



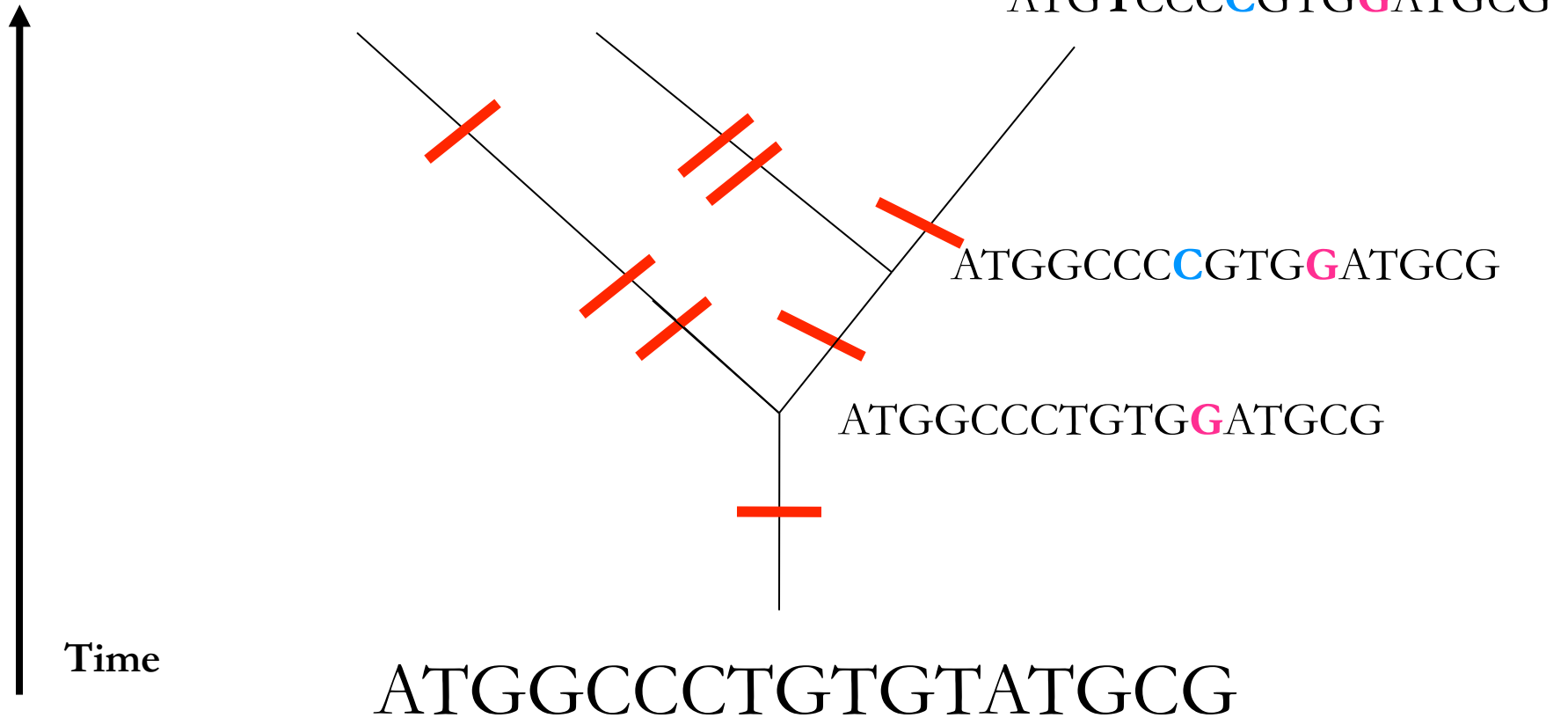
A history of mutations



A history of mutations



A history of mutations



A history of mutations

ATGGCAATGTG**G**ATGCA

ATGGCCCC**C**GTG**G**AACCG

ATGTCCCC**C**GTG**G**ATGCG

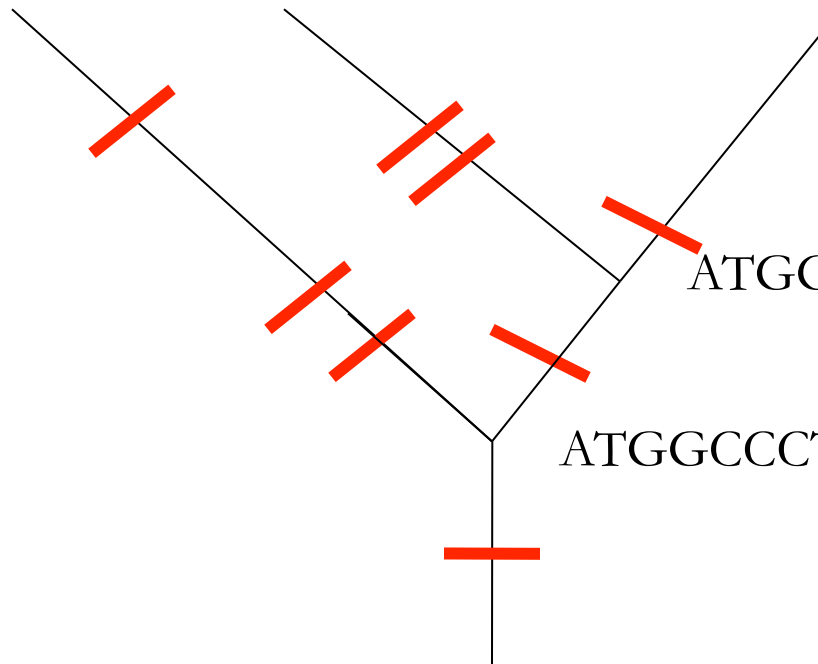
ATGGCCCC**C**GTG**G**ATGCG

ATGGCCCTGTG**G**ATGCG

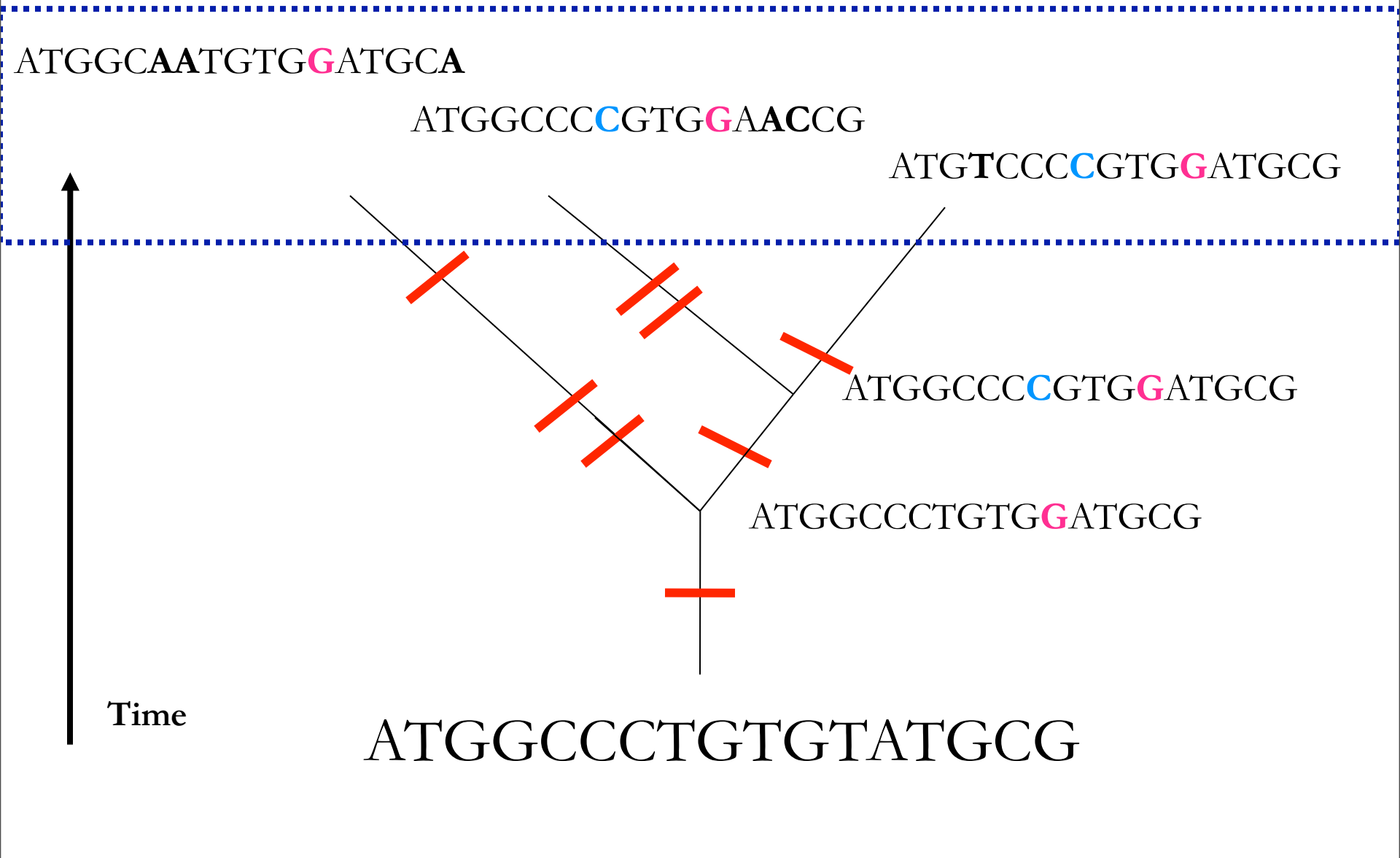
ATGGCCCTGTGTATGCG



Time

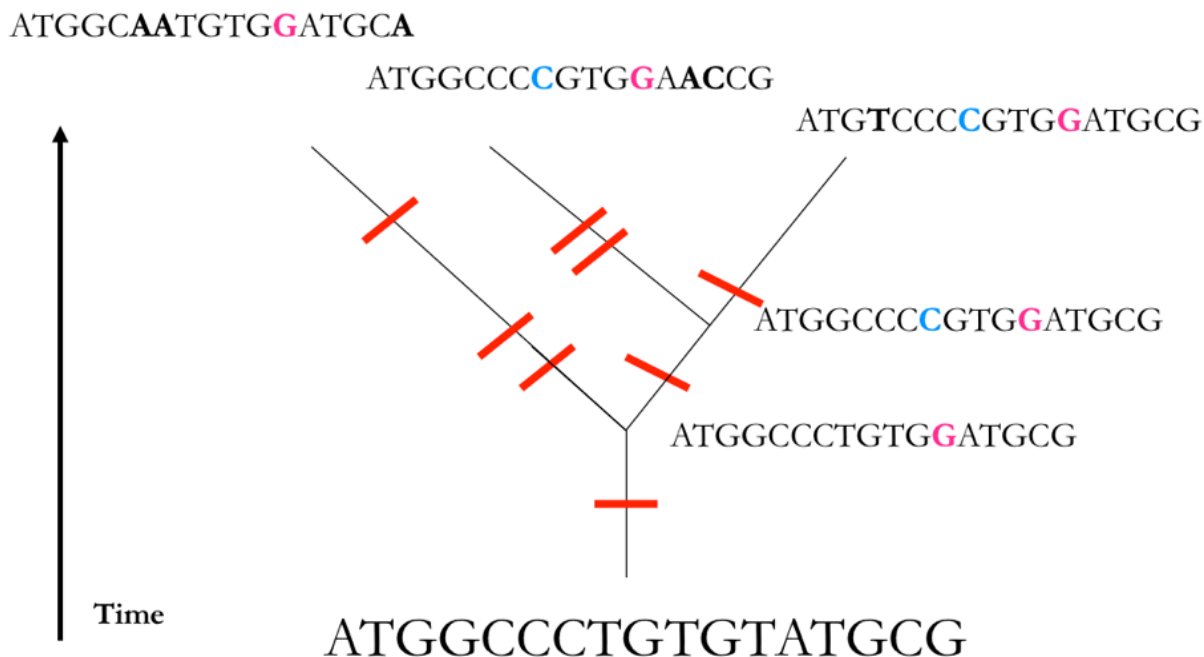


A history of mutations

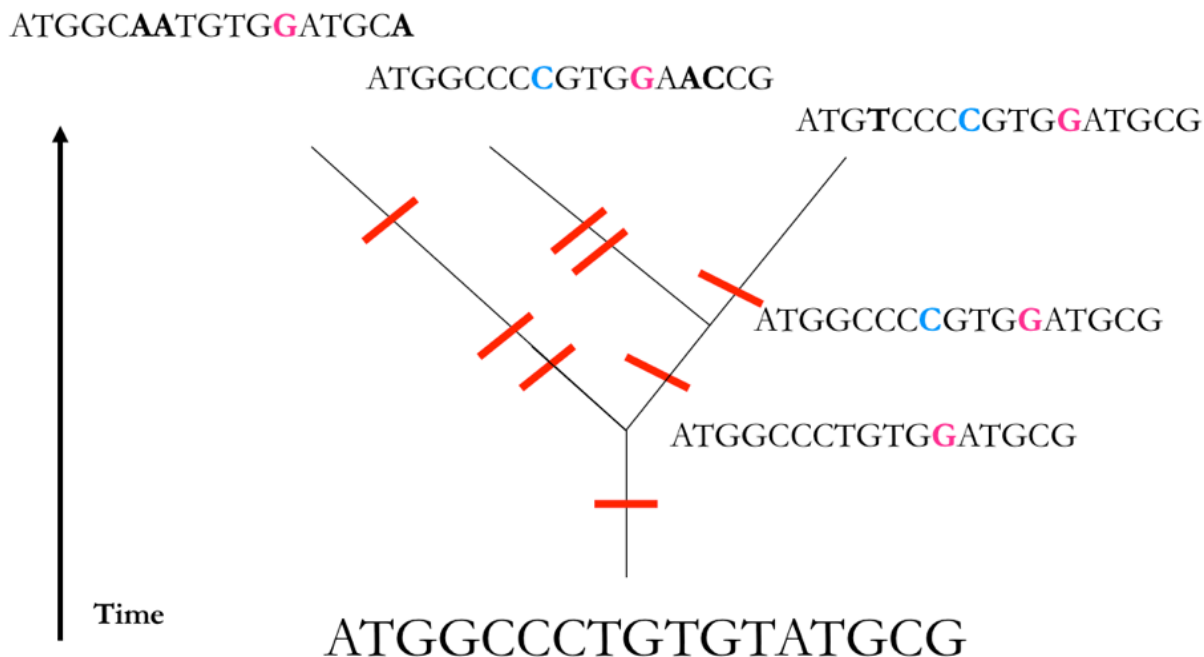


“DNA alignment”

- Species1: ATGGC**AA**TGTG**G**ATGCA**A**
- Species2: ATGGCCCC**C**GTG**G**AA**AC**CG
- Species3: ATG**T**CCCC**C**GTG**G**ATGCG



- Species1: ATGGC**AA**TGTG**G**ATGCA**A**
- Species2: ATGGCCCC**C**GTG**G**AA**AC**CG
- Species3: ATG**T**CCCC**C**GTG**G**ATGCG



- Species1: ATGGC**AA**TGTG**G**ATGCA
 - Species2: ATGGCCCC**C**GTG**G**A**AC**CG
 - Species3: ATG**T**CCCC**C**GTG**G**ATGCG
- } 6
} 3

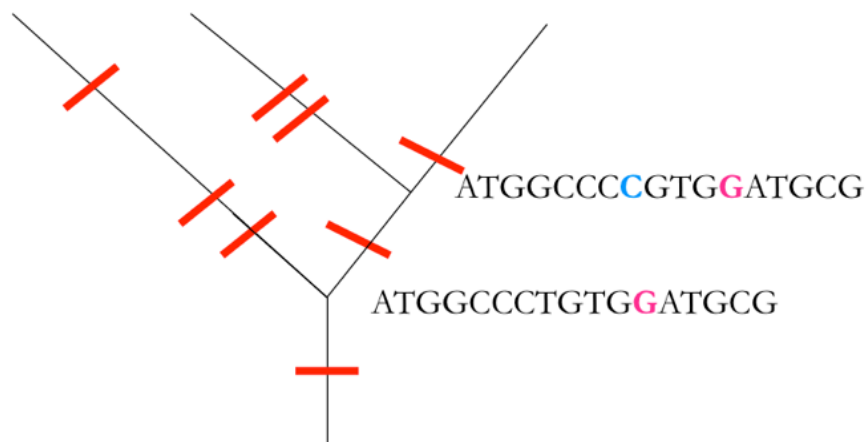
ATGGCA**AA**TGTG**G**ATGCA

ATGGCCCC**C**GTG**G**A**AC**CG

ATG**T**CCCC**C**GTG**G**ATGCG



Time



ATGGCCCTGTG**G**ATGCG

“DNA alignment”

- Species1: ATGGC**A**ATGTG**G**ATGCA
 - Species2: ATGGCCC**C**GTG**G**A**A**CCG
 - Species3: ATG**T**CCCC**C**GTG**G**ATGCG
- $\left. \begin{array}{l} 6 \\ 3 \end{array} \right\} 5$

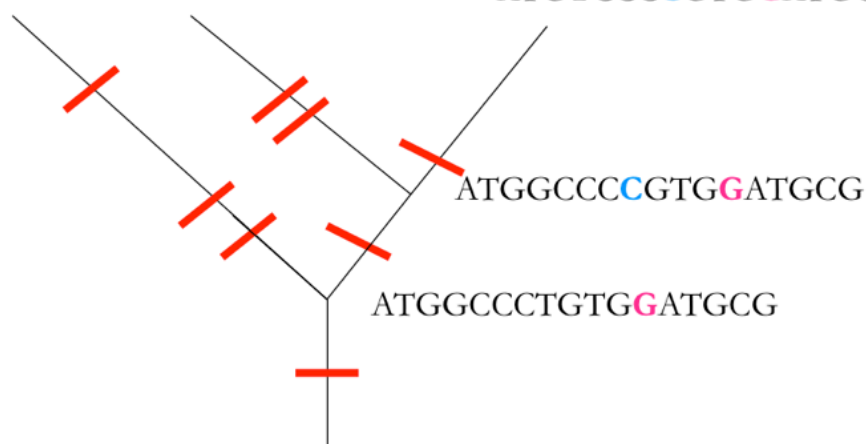
ATGGCAATGTG**G**ATGCA

ATGGCCC**C**GTG**G**A**A**CCG

ATG**T**CCCC**C**GTG**G**ATGCG



Time



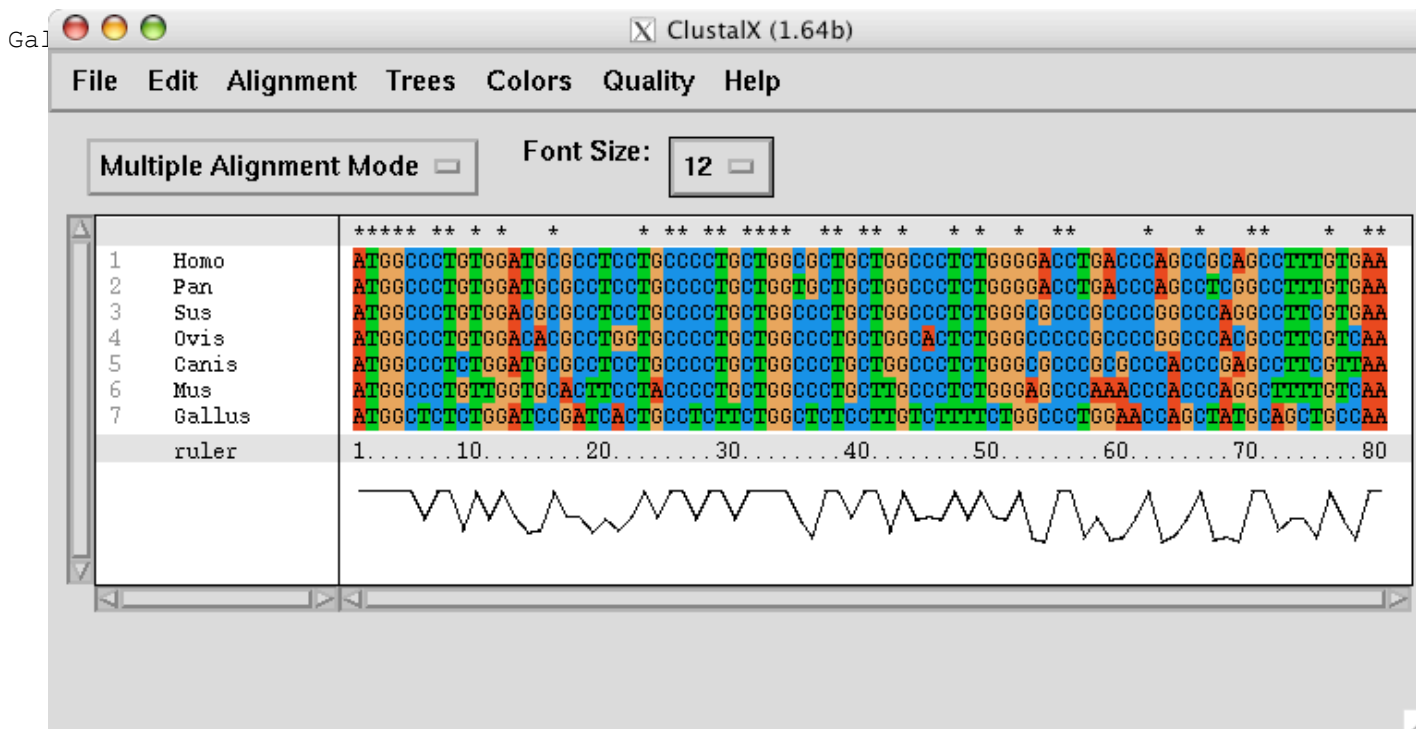
ATGGCCCTGTG**G**ATGCG

- ## Insulin from 7 different species

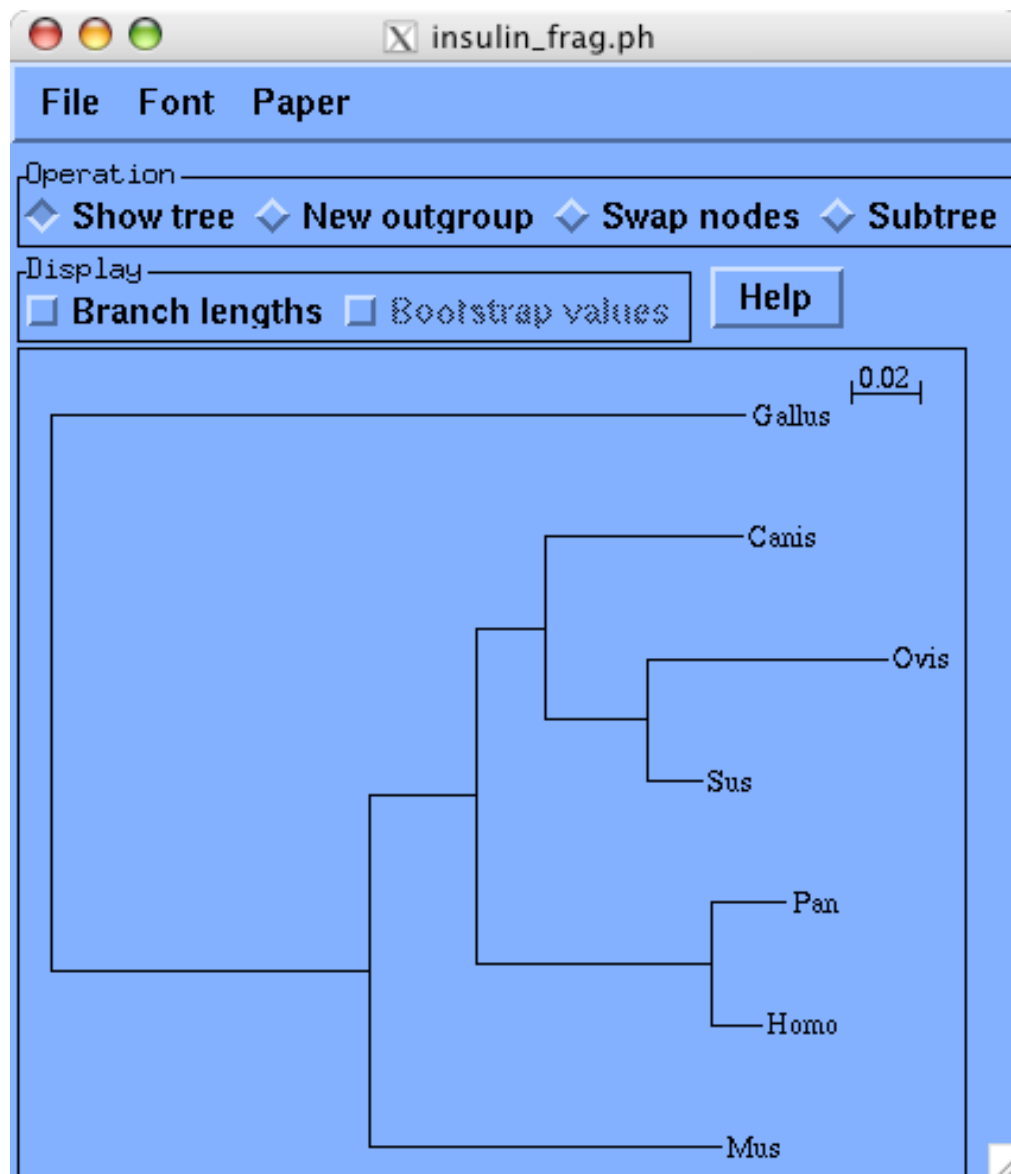
- Homo: ATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGAGCCTTTGTGAA
- Pan: ATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGTGCTGCTGGCCCTCTGGGGACCTGACCCAGCCTCGGCCTTTGTGAA
- Sus: ATGGCCCTGTGGACGCGCCTCCTGCCCCTGCTGGCCCTGCTGGCCCTCTGGGGCGCCCGCCCCGGCCCCAGGCCTTCGTGAA
- Ovis: ATGGCCCTGTGGACACGCCTGGTGCCCCTGCTGGCCCTGCTGGCACTCTGGGGCCCCCGCCCCGGCCCCACGCCTTCGTCAA
- Canis: ATGGCCCTCTGGATGCGCCTCCTGCCCCTGCTGGCCCTGCTGGCCCTCTGGGGCGCCCGCGCCACCCGAGCCTTCGTAA
- Mus: ATGGCCCTGTTGGTGCACTTCCTACCCCTGCTGGCCCTGCTTGCCCTCTGGGAGCCCAAACCCACCCAGGCTTTGTCAA
- Gallus: ATGGCTCTCTGGATCCGATCACTGCCTCTTCTGGCTCTCCTTGTCTTTTCTGGCCCTGGAACCAGCTATGCAGCTGCCAA

• Insulin from 7 different species

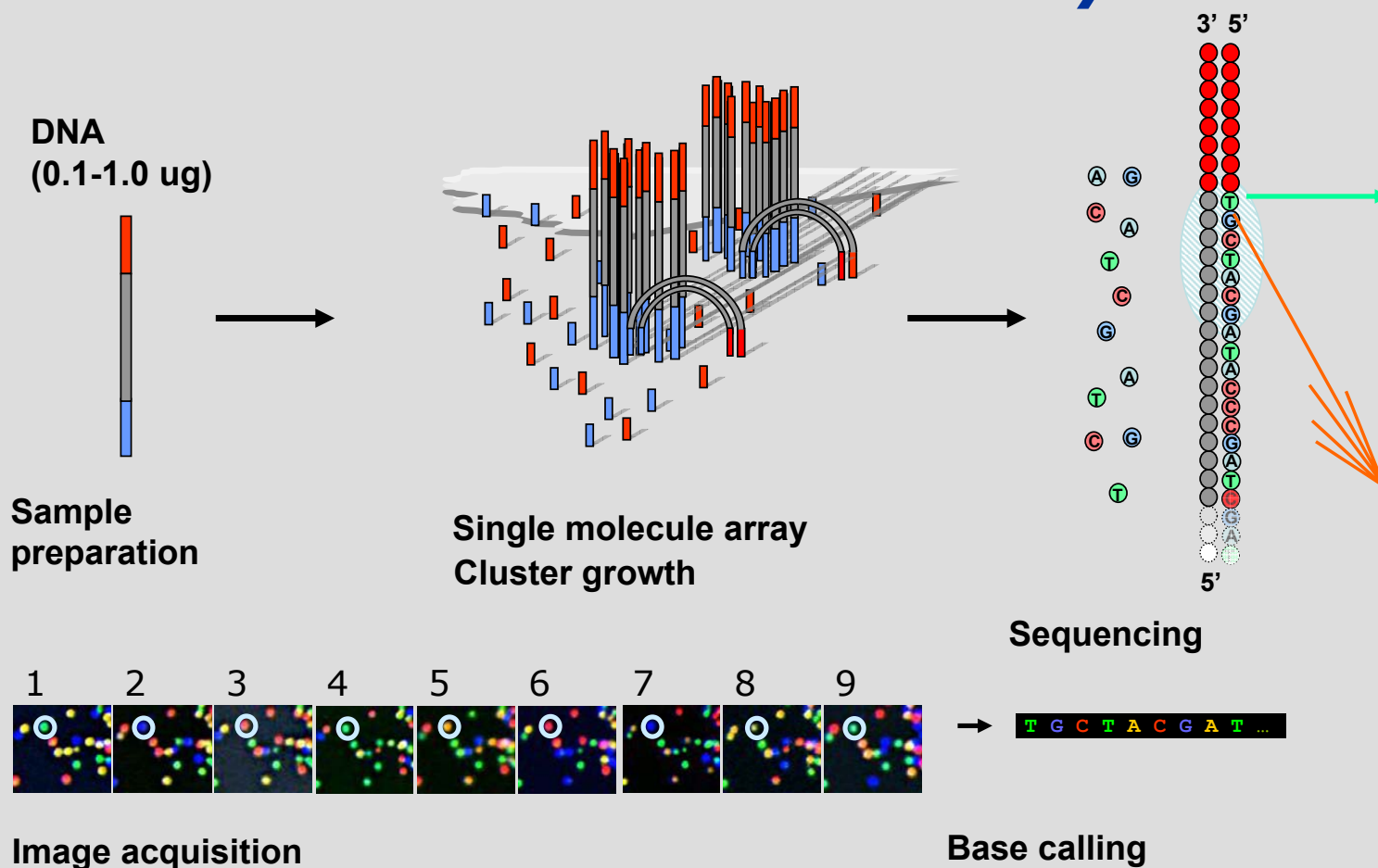
- Homo: ATGGCCCTGTGGATGCGCCTCCTGCCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGACGCTTTGTGAA
- Pan: ATGGCCCTGTGGATGCGCCTCCTGCCCCCTGCTGGTGCTGCTGGCCCTCTGGGGACCTGACCCAGCCTCGGCCTTTGTGAA
- Sus: ATGGCCCTGTGGACGCGCCTCCTGCCCCCTGCTGGCCCTGCTGGCCCTCTGGGGCGCCCGCCCCGGCCAGGCCTTCGTGAA
- Ovis: ATGGCCCTGTGGACACGCCTGGTGCCCCCTGCTGGCCCTGCTGGCACTCTGGGCCCCCGCCCCGGCCACGCCTTCGTCAA
- Canis: ATGGCCCTCTGGATGCGCCTCCTGCCCCCTGCTGGCCCTGCTGGCCCTCTGGGGCGCCCGCGCCACCCGAGCCTTCGTTAA
- Mus: ATGGCCCTGTTGGTGCACTTCCTACCCCTGCTGGCCCTGCTTGCCCTCTGGGAGCCCAAACCCACCCAGGCTTTTGTCAA
- Gal



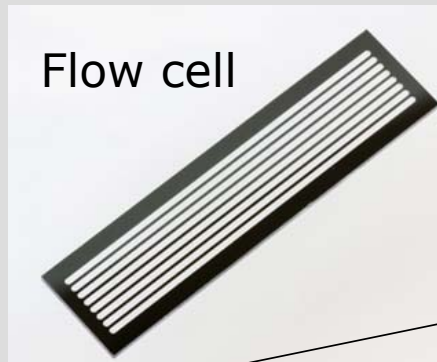
Real life example: Tree



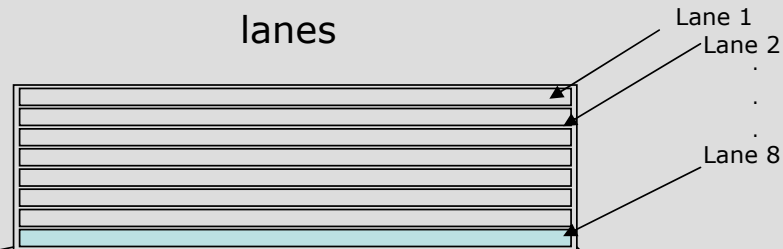
Illumina Sequencing Technology *Reversible Terminator Chemistry*



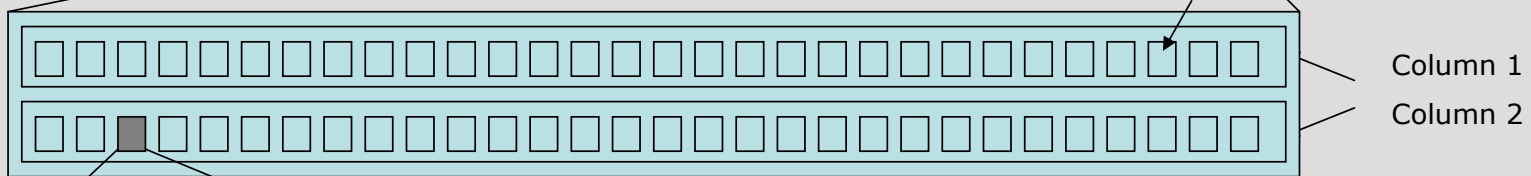
Illumina Sequencing Technology



A **flow cell** contains eight lanes



Each **lane** contains **two columns** of tiles

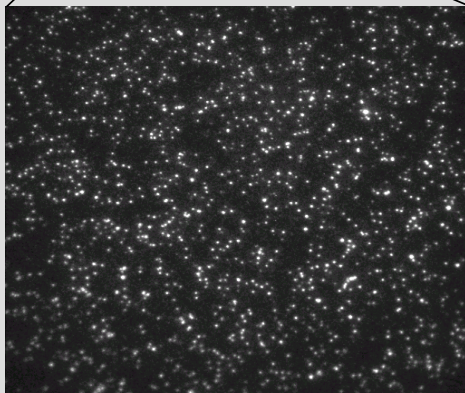


Each **column** contains **multiple tiles** – total 120
Each tile is imaged four times per cycle – one image per base.

~340.000 clusters/tile ->

~40.000.000 clusters/lane ->

~320.000.000 clusters/flowcell

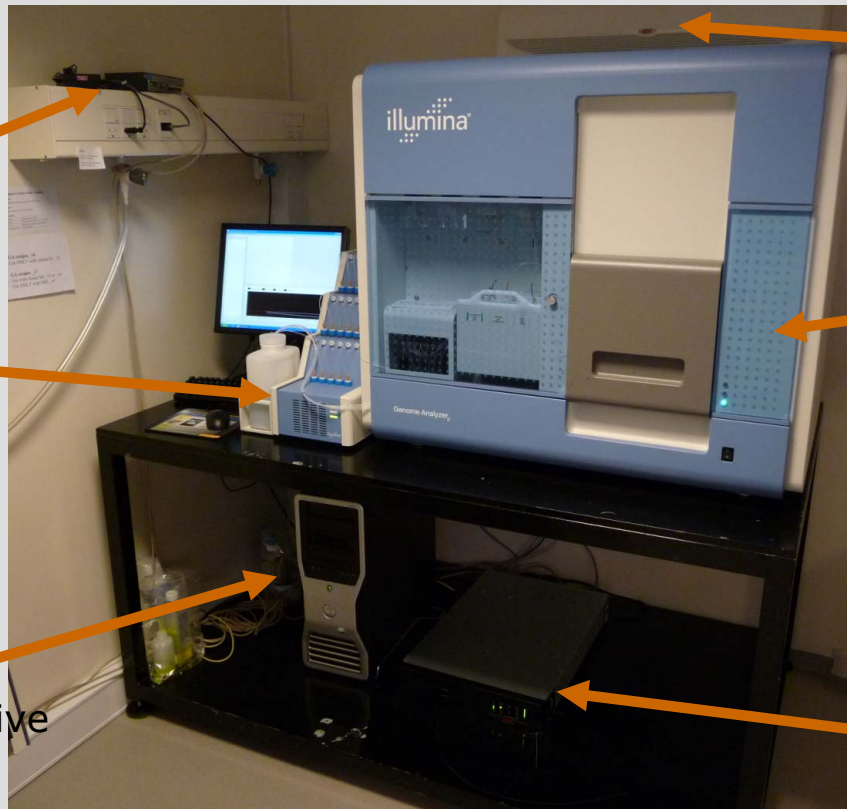


Illumina GA Sequencing technology

Switch and 100 Mbps network to pipeline computer

Paired End (PE) module

GA PC
 • 2.66 GHz cpu
 • 3 GB RAM
 • 80 GB hard drive



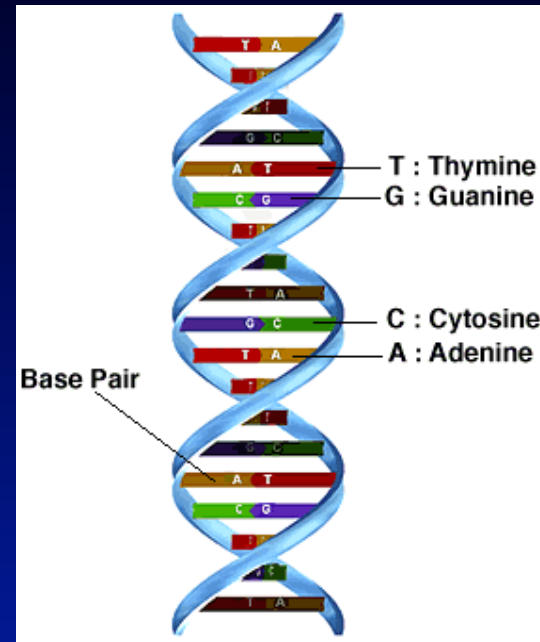
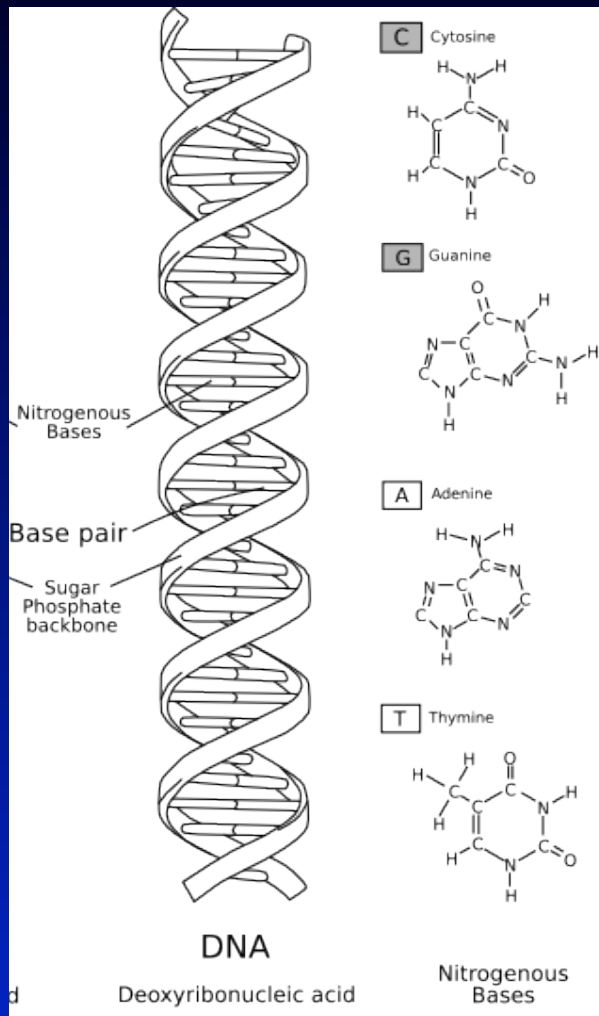
Cooling unit!

Genome Analyzer (GAIIx)

Uninterruptible Power Supply (UPS)
 • Back up for ~10 min

Illumina Genome Analyzer IIx, "Oban"

Symbolic representation of DNA structure

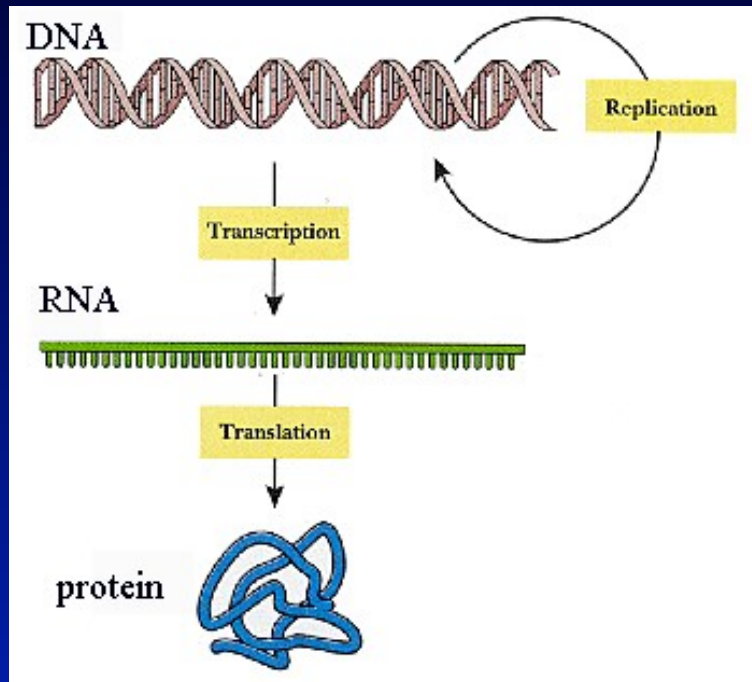


- DNA molecule is a linear polymer
- Structure can be represented as string of 4 symbols: ACTG
- These “sequences” can be analyzed mathematically/linguistically

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS

sdag den 20. september 2011

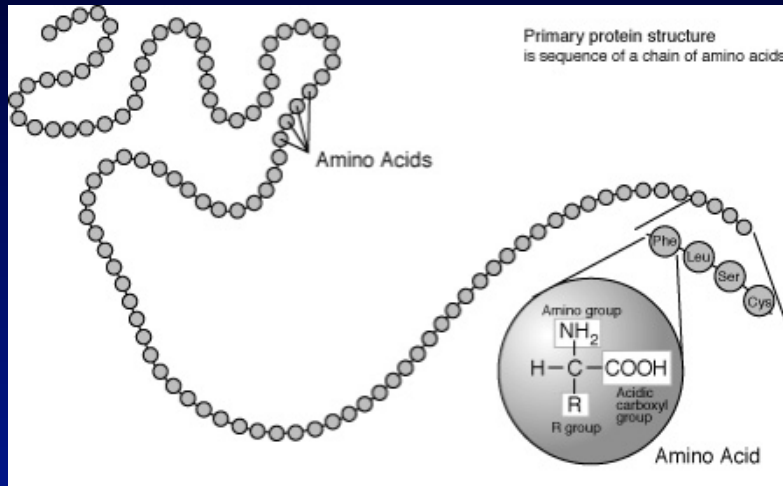
DNA --> RNA --> protein



Standard Genetic Code

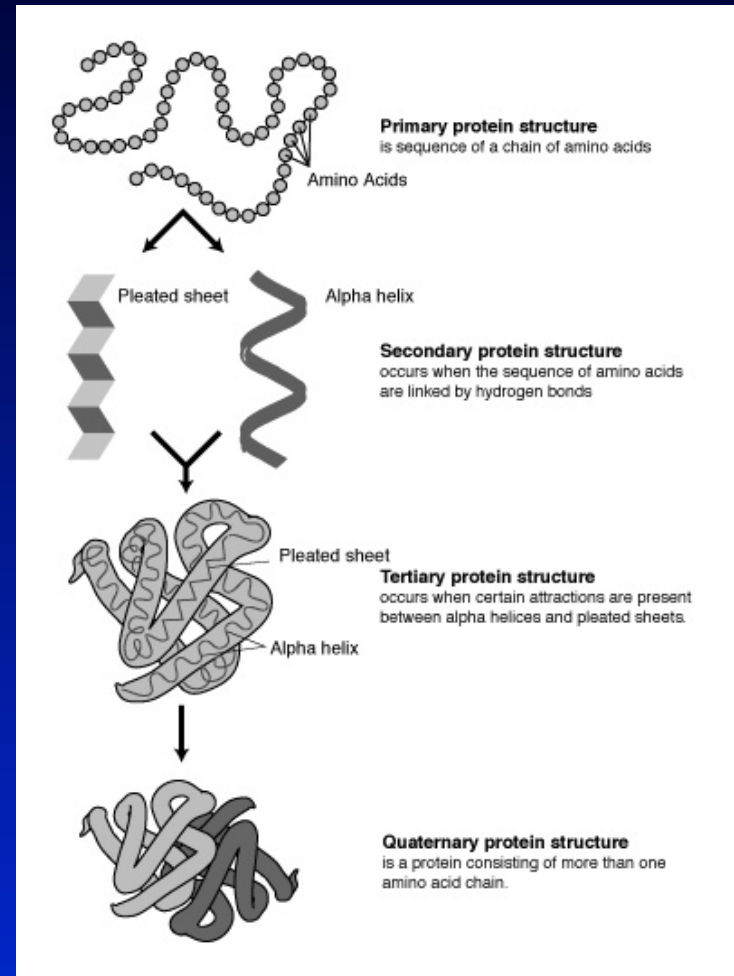
	T			C			A			G			
T	TTT	Phe	F	TCT	Ser	S	TAT	Tyr	Y	TGT	Cys	C	T
	TTC	Phe	F	TCC	Ser	S	TAC	Tyr	Y	TGC	Cys	C	C
	TTA	Leu	L	TCA	Ser	S	TAA	Och *		TGA	Opa *		A
	TTG	Leu	L	TCG	Ser	S	TAG	Amb *		TGG	Trp	W	G
C	CTT	Leu	L	CCT	Pro	P	CAT	His	H	CGT	Arg	R	T
	CTC	Leu	L	CCC	Pro	P	CAC	His	H	CGC	Arg	R	C
	CTA	Leu	L	CCA	Pro	P	CAA	Gln	Q	CGA	Arg	R	A
	CTG	Leu	L	CCG	Pro	P	CAG	Gln	Q	CGG	Arg	R	G
A	ATT	Ile	I	ACT	Thr	T	AAT	Asn	N	AGT	Ser	S	T
	ATC	Ile	I	ACC	Thr	T	AAC	Asn	N	AGC	Ser	S	C
	ATA	Ile	I	ACA	Thr	T	AAA	Lys	K	AGA	Arg	R	A
	ATG	Met	M	ACG	Thr	T	AAG	Lys	K	AGG	Arg	R	G
G	GTT	Val	V	GCT	Ala	A	GAT	Asp	D	GGT	Gly	G	T
	GTC	Val	V	GCC	Ala	A	GAC	Asp	D	GGC	Gly	G	C
	GTA	Val	V	GCA	Ala	A	GAA	Glu	E	GGA	Gly	G	A
	GTG	Val	V	GCG	Ala	A	GAG	Glu	E	GGG	Gly	G	G

Symbolic representation of protein structure



- Proteins are linear polymers
- Built from 20 amino acids
- Can be represented as string of 20 symbols

ACDEFGHIKLMNPQRSTVWY



NCBI databases

The screenshot shows the NCBI website in a web browser. The browser's address bar displays <http://www.ncbi.nlm.nih.gov/>. The page features a navigation bar with links to NCBI, Resources, and How To. A search bar is located at the top right, with the text "human globin" entered. The main content area is divided into several sections: "Resources" on the left, "Welcome to NCBI" in the center, "Genome Reference Consortium" below it, "How To..." with a list of tasks, "NLN/NCBI H1N1 Flu Resources" at the bottom left, "Popular Resources" on the right, and "NCBI News" at the bottom right. The "Resources" section includes links to NCBI Home, All Resources (A-Z), Literature, DNA & RNA, Proteins, Sequence Analysis, Genes & Expression, Genomes, Maps & Markers, Domains & Structures, Genetics & Medicine, Taxonomy, Data & Software, Training & Tutorials, Homology, Small Molecules, and Variation. The "Welcome to NCBI" section provides a brief overview of the center's mission and links to more information. The "Genome Reference Consortium" section describes its goal to improve human and mouse reference assemblies. The "How To..." section lists tasks such as obtaining full text, retrieving sequences, finding homologs, and designing PCR primers. The "Popular Resources" section lists links to PubMed, PubMed Central, Bookshelf, BLAST, Gene, Nucleotide, Protein, GEO, Conserved Domains, Structure, and PubChem. The "NCBI News" section features two news items: "November and October News" dated 02 Dec 2009 and "NCBI News - September 2009" dated 05 Oct 2009.

National Center for Biotechnology Information

Search

Resources

- NCBI Home
- All Resources (A-Z)
- Literature
- DNA & RNA
- Proteins
- Sequence Analysis
- Genes & Expression
- Genomes
- Maps & Markers
- Domains & Structures
- Genetics & Medicine
- Taxonomy
- Data & Software
- Training & Tutorials
- Homology
- Small Molecules
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[More about the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS](#)

Genome Reference Consortium

Formed to improve human and mouse reference assemblies, GRC will fix loci misrepresented in reference assembly, fill remaining gaps, and make alternate representations of complex loci.

How To...

- Obtain the full text of an article
- Retrieve all sequences for an organism or taxon
- Find a homolog for a gene in another organism
- Find genes associated with a phenotype or disease
- Design PCR primers and check them for specificity
- Find the function of a gene or gene product
- Determine conserved synteny between the genomes of two organisms

[See all ...](#)

NLM/NCBI H1N1 Flu Resources

Popular Resources

- PubMed
- PubMed Central
- Bookshelf
- BLAST
- Gene
- Nucleotide
- Protein
- GEO
- Conserved Domains
- Structure
- PubChem

NCBI News

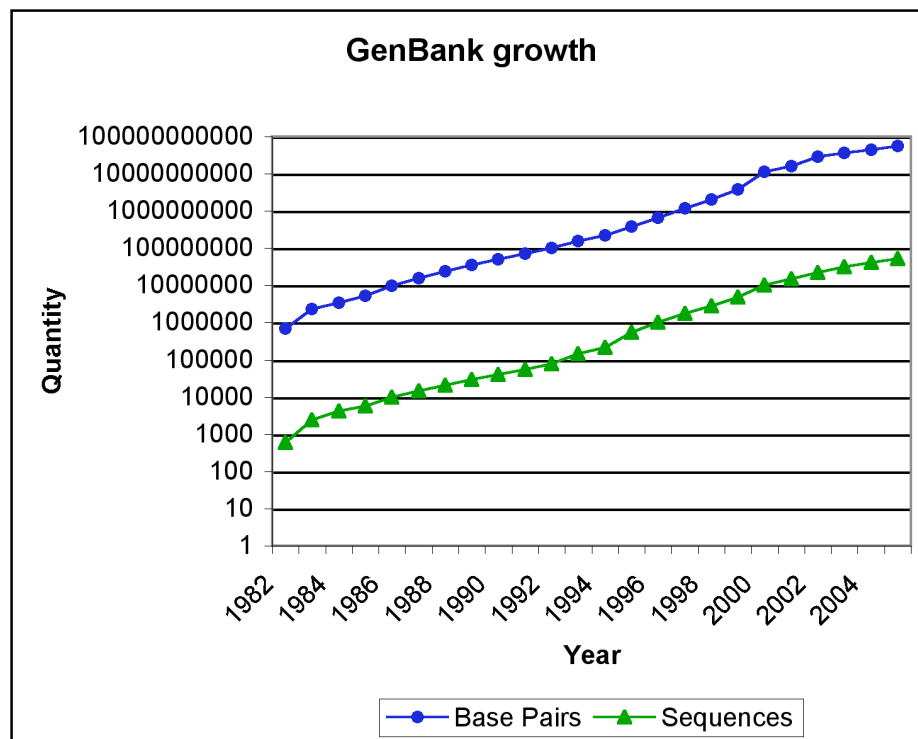
November and October News 02 Dec 2009
Featured: New Discovery-oriented PubMed and NCBI Homepage. T...

NCBI News - September 2009 05 Oct 2009
The September 2009 issue of the NCBI News is available ...

NCBI News - August 2009 19 Aug 2009
The August 2009 issue of the NCBI News is available online. ...

[More...](#)

- GenBank is one of the main international DNA databases.
- GenBank is hosted by NCBI: *National Center for Biotechnology Information.*
- GenBank has existed since 1982.
- The database is public - no restrictions on the use of the data within.



GenBank format

[illegible]

Header

Indeholder
information ang.
Organisme,
publikation,
Accession ID mm.

FEATURE blok

Indeholder en beskrivelse af forskellige elementer i DNA sekvensen.

CDS: **C**oding **S**equ^{ence}. Indeholder koordinater på den protein kodende del af et gen. Bemærk de tre intervaller.

ORIGIN blok

Indeholder selve
DNA sekvensen.

Introduction 188 [info](#) [contact](#) [help](#) [version 2765](#)

- Originates from the GenBank database.
- Contains both a DNA sequence and annotation of feature (e.g. Location of genes).

(handout)

GenBank format - HEADER

LOCUS CMGLOAD 1185 bp DNA linear VRT 18-APR-2005
 DEFINITION Cairina moschata (duck) gene for alpha-D globin.
 ACCESSION X01831
 VERSION X01831.1 GI:62724
 KEYWORDS alpha-globin; globin.
 SOURCE Cairina moschata (Muscovy duck)
 ORGANISM Cairina moschata
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
 Archosauria; Aves; Neognathae; Anseriformes; Anatidae; Cairina.
 REFERENCE 1 (bases 1 to 1185)
 AUTHORS Erbil,C. and Niessing,J.
 TITLE The primary structure of the duck alpha D-globin gene: an unusual
 5' splice junction sequence
 JOURNAL EMBO J. 2 (8), 1339-1343 (1983)
 PUBMED 10872328
 COMMENT Data kindly reviewed (13-NOV-1985) by J. Niessing.

GenBank format - ORIGIN section

ORIGIN

```

1  ctgcgtggcc tcagcccctc caccctcca cgctgataag ataaggccag ggcgggagcg
61  cagggtgcta taagagctcg gcccgcggg tgtctccacc acagaaacc gtcagttgcc
121 agcctgccac gccgctgccg ccatgctgac cgccgaggac aagaagctca tcgtgcaggt
181 gtgggagaag gtggctggcc accaggagga attcggaagt gaagctctgc agaggtgtgg
241 gctgggcca gggggcactc acaggggtgg cagcaggag caggagccct gcagcgggtg
301 tgggctggga ccagagcgc cacggggtgc gggctgagat gggcaaagca gcagggcacc
361 aaaactgact ggctcgctc cggcaggatg ttctcgcct acccccagac caagacctac
421 tccccccact tcgacctgca tcccggtctt gaacagggtc gtggccatgg caagaaagtg
481 gcggctgccc tgggcaatgc cgtgaagagc ctggacaacc tcagccaggc cctgtctgag
541 ctcagcaacc tgcatgccta caacctgcgt gttgacctg tcaacttcaa ggcaagcggg
601 gactagggtc cttgggtctg ggggtctgag ggtgtggggt gcaggggtctg ggggtccagg
661 ggtctgagtt tcctggggtc tggcagtcct gggggctgag ggccagggtc ctgtggtctt
721 gggtagcagg gtcctggggg ccagcagcca gacagcagg gctgggattg catctgggat
781 gtgggcccaga ggctgggatt gtgtttggaa tgggagctgg gcaggggcta gggccagggt
841 gggggactca gggcctcagg gggactcggg gggggactga gggagactca gggccatctg
901 tccggagcag gggtagtaag ccctggtttg ccttgagct gctggcacag tgcttccagg
961 tgggtgctggc cgcacacctg ggcaaagact acagccccga gatgcatgct gcctttgaca
1021 agttcttgct cgccgtggct gccgtgctgg ctgaaaagta cagatgagcc actgcctgca
1081 cccttgacc ttcaataaag acaccattac cacagctctg tgtctgtgtg tgctgggact
1141 gggcatcggg ggtcccaggg agggctgggt tgcttcaca catcc

```

//

GenBank format - FEATURE section

```

FEATURES                     Location/Qualifiers
    source                    1..1185
                              /organism="Cairina moschata"
                              /mol_type="genomic DNA"
                              /db_xref="taxon:8855"
    CAAT_signal               20..24
    TATA_signal               69..73
    precursor_RNA             101..1114
                              /note="primary transcript"
    exon                      101..234
                              /number=1
    CDS                       join(143..234,387..591,939..1067)
                              /codon_start=1
                              /product="alpha D-globin"
                              /protein_id="CAA25966.2"
                              /db_xref="GI:4455876"
                              /db_xref="GOA:P02003"
                              /db_xref="InterPro:IPR000971"
                              /db_xref="InterPro:IPR002338"
                              /db_xref="InterPro:IPR002340"
                              /db_xref="InterPro:IPR009050"
                              /db_xref="UniProt/Swiss-Prot:P02003"
                              /translation="MLTAEDKKLIVQVWEKVAGHQEEFGSEALQRMFLAYPQTKTYFP
HFDLHPGSEQVRGHGKKVAAALGNAVKSLDNLSQALSELSNLHAYNLRVDPVNFKLLA
QCFQVVLAAHLGKDYSPMHAAFDKFLSAVAAVLAEKYR"
    repeat_region             227..246
                              /note="direct repeat 1"
    intron                    235..386
                              /number=1
    repeat_region             289..309
                              /note="direct repeat 1"
    exon                      387..591
                              /number=2
    intron                    592..939
                              /number=2
    exon                      940..1114
                              /number=3
    polyA_signal              1095..1100
    polyA_signal              1114
  
```

NCBI databases: fasta format

The screenshot displays the NCBI Nucleotide database interface. The browser address bar shows the URL: [http://www.ncbi.nlm.nih.gov/nuccore/28302130?report=fasta&log\\$=seqview&from=54&to=497](http://www.ncbi.nlm.nih.gov/nuccore/28302130?report=fasta&log$=seqview&from=54&to=497). The page title is "Nucleotide - Homo sapiens hemoglobin, gamma A (HBG1), mRNA". The NCBI logo is visible in the top left. The search bar contains "Nucleotide" and "for". The format is set to "FASTA". The sequence is displayed in FASTA format, starting with >gi|28302130:54-497 Homo sapiens hemoglobin, gamma A (HBG1), mRNA. The sequence is shown in a single line, wrapped at 60 characters per line. The sequence is: ATGGGTCATTTCACAGAGGAGGACAAGGCTACTATCACAGCCTGTGGGGCAAGGTGAATGTGGAAGATGCTGGAGGAGAAACCTGGGAAGGCTCCTGGTTGTCTACCCATGGACCCAGAGGTTCTTTGACAGCTTTGGCAACCTGTCTCTGCTCTGCCATCATGGGCAACCCCAAGTCAAGGCACATGGCAAGAAGGTGCTGACTTCCTTGGGAGATGCCACAAAGCACCTGGATGATCTCAAGGGCACCTTTGCCAGCTGAGTGAAGTGCACCTGTGACAAGCTGCATGTGGATCCTGAGAACTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTTGGCAATCCATTTCGGCAAGAATTACCCCTGAGGTGCAGGCTTCCTGGCAGAAGATGGTGACTGCAGTGGCCAGTGCCCTGCTCTCCAGATACCCTGA.

Showing 444 bp region from base 54 to 497.

NCBI Reference Sequence: NM_000559.2

Homo sapiens hemoglobin, gamma A (HBG1), mRNA

>gi|28302130:54-497 Homo sapiens hemoglobin, gamma A (HBG1), mRNA
ATGGGTCATTTCACAGAGGAGGACAAGGCTACTATCACAGCCTGTGGGGCAAGGTGAATGTGGAAGATGCTGGAGGAGAAACCTGGGAAGGCTCCTGGTTGTCTACCCATGGACCCAGAGGTTCTTTGACAGCTTTGGCAACCTGTCTCTGCTCTGCCATCATGGGCAACCCCAAGTCAAGGCACATGGCAAGAAGGTGCTGACTTCCTTGGGAGATGCCACAAAGCACCTGGATGATCTCAAGGGCACCTTTGCCAGCTGAGTGAAGTGCACCTGTGACAAGCTGCATGTGGATCCTGAGAACTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTTGGCAATCCATTTCGGCAAGAATTACCCCTGAGGTGCAGGCTTCCTGGCAGAAGATGGTGACTGCAGTGGCCAGTGCCCTGCTCTCCAGATACCCTGA

Change Region Shown

☐ Whole sequence
☒ Selected Region
from: 54 to: 497
[Update View](#)

Customize View

Analyze This Sequence

- ▶ Run BLAST
- ▶ Pick Primers

Articles about the HBG1 gene

- ▶ Molecular analysis of gamma-globin promoters, HS-111 and [Hemoglobin. 2009]
- ▶ A genome-wide association identified the common genetic variant [Hum Genet. 2009]
- ▶ Expression of miR-210 during erythroid differentiation and induction [BMB Rep. 2009]

» See all...

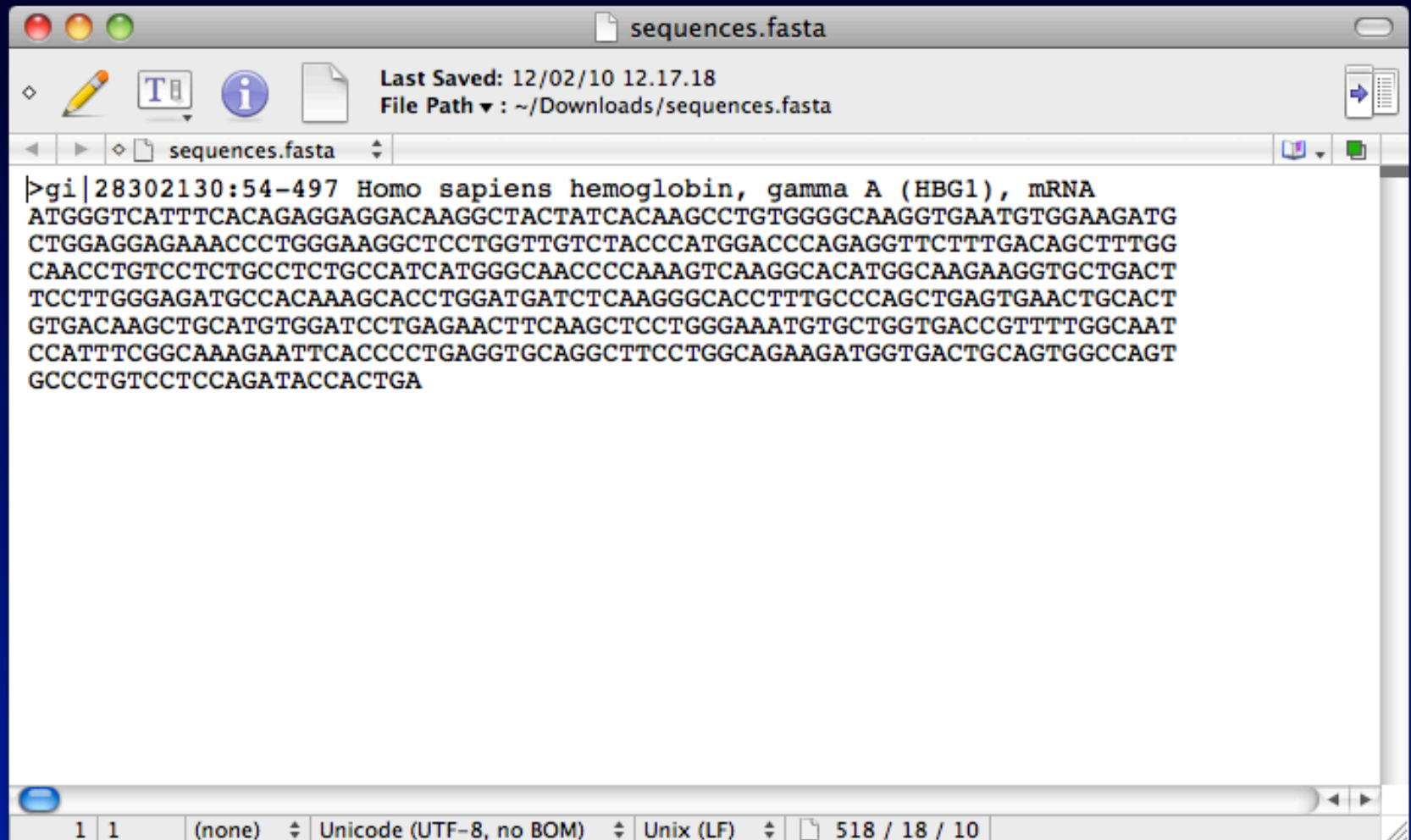
RefSeq Protein Product

See the reference protein sequence for A-gamma globin (NP_000550.2).

More about the HBG1 gene

The gamma-globin genes (HBG1 and HBG2)

FASTA file



```
>gi|28302130:54-497 Homo sapiens hemoglobin, gamma A (HBG1), mRNA
ATGGGTCATTTACAGAGGAGGACAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAATGTGGAAGATG
CTGGAGGAGAAACCCTGGGAAGGCTCCTGGTTGTCTACCCATGGACCCAGAGGTTCTTTGACAGCTTTGG
CAACCTGTCCTCTGCCTCTGCCATCATGGGCAACCCCAAAGTCAAGGCACATGGCAAGAAGGTGCTGACT
TCCTTGGGAGATGCCACAAAGCACCTGGATGATCTCAAGGGCACCTTTGCCCAGCTGAGTGAAGTGCAGT
GTGACAAGCTGCATGTGGATCCTGAGAACTTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTTGGCAAT
CCATTTTCGGCAAAGAATTCACCCCTGAGGTGCAGGCTTCCTGGCAGAAGATGGTGACTGCAGTGGCCAGT
GCCCTGTCCTCCAGATAACCACTGA
```